Charles K. Chui
Qingtang Jiang

Applied Mathematics

Data Compression, Spectral Methods,
Fourier Analysis, Wavelets,
and Applications

Mathematics Textbooks for Science and Engineering

Volume 2

Charles K. Chui · Qingtang Jiang

Applied Mathematics

Data Compression, Spectral Methods, Fourier
Analysis, Wavelets, and Applications

Charles K. Chui
Department of Statistics
Stanford University
Stanford, CA
USA

Qingtang Jiang
Department of Mathematics
  and Computer Science
University of Missouri
St. Louis, MO
USA

Printed on acid-free paper

Series Information

Textbooks in the series 'Mathematics Textbooks for Science and Engineering' will be aimed at the broad mathematics, science and engineering undergraduate and graduate levels, covering all areas of applied and applicable mathematics, interpreted in the broadest sense.

Series Editor

Charles K. Chui
Stanford University, Stanford, CA, USA

Atlantis Press
8 square des Bouleaux
75019 Paris, France

For more information on this series and our other book series, please visit our website www.atlantis-press.com

Editorial

Recent years have witnessed an extraordinarily rapid advance in the direction of information technology within both the scientific and engineering disciplines. In addition, the current profound technological advances of data acquisition devices and transmission systems contribute enormously to the continuing exponential growth of data information that requires much better data processing tools. To meet such urgent demands, innovative mathematical theory, methods, and algorithms must be developed, with emphasis on such application areas as complex data organization, contaminated noise removal, corrupted data repair, lost data recovery, reduction of data volume, data dimensionality reduction, data compression, data understanding and visualization, as well as data security and encryption.

The revolution of the data information explosion as mentioned above demands early mathematical training with emphasis on data manipulation at the college level and beyond. The Atlantis book series, "Mathematics Textbooks for Science and Engineering (MTSE)", is founded to meet the needs of such mathematics textbooks that can be used for both classroom teaching and self-study. For the benefit of students and readers from the interdisciplinary areas of mathematics, computer science, physical and biological sciences, and various engineering specialties, contributing authors are requested to keep in mind that the writings for the MTSE book series should be elementary and relatively easy to read, with sufficient examples and exercises. We welcome submission of such book manuscripts from all who agree with us on this point of view.

This second volume is intended to be a comprehensive textbook in "Contemporary Applied Mathematics", with emphasis in the following five areas: spectral methods with applications to data analysis and dimensionality reduction; Fourier theory and methods with applications to time-frequency analysis and solution of partial differential equations; Wavelet time-scale analysis and methods, with in-depth study of the lifting schemes, wavelet regularity theory, and convergence of cascade algorithms; Computational algorithms, including fast Fourier transform, fast cosine transform, and lapped transform; and Information theory with applications to image and video compression. This book is self-contained, with writing

style friendly towards the teacher and reader. It is intended to be a textbook, suitable for teaching in a variety of courses, including: Applied Mathematics, Applied Linear Algebra, Applied Fourier Analysis, Wavelet Analysis, and Engineering Mathematics, both at the undergraduate and beginning graduate levels.

Charles K. Chui
Menlo Park, CA

Preface

Mathematics was coined the "queen of science" by the "prince of mathematicians," Carl Friedrich Gauss, one of the greatest mathematicians of all time. Indeed, the name of Gauss is associated with essentially all areas of mathematics. It is therefore safe to assume that to Gauss there was no clear boundary between "pure mathematics" and "applied mathematics". To ensure financial independence, Gauss decided on a stable career in astronomy, which is one of the oldest sciences and was perhaps the most popular one during the eighteenth and nineteenth centuries. In his study of celestial motion and orbits and a diversity of disciplines later in his career, including (in chronological order): geodesy, magnetism, dioptrics, and actuarial science, Gauss has developed a vast volume of mathematical methods and tools that are still instrumental to our current study of applied mathematics.

During the twentieth century, with the exciting development of quantum field theory, with the prosperity of the aviation industry, and with the bullish activity in financial market trading, and so forth, much attention was paid to the mathematical research and development in the general discipline of partial differential equations (PDEs). Indeed, the non-relativistic modeling of quantum mechanics is described by the Schrödinger equation; the fluid flow formulation, as an extension of Newtonian physics by incorporating motion and stress, is modeled by the Navier-Stokes equation; and option stock trading with minimum risk can be modeled by the Black-Scholes equation. All of these equations are PDEs. In general, PDEs are used to describe a wide variety of phenomena, including: heat diffusion, sound wave propagation, electromagnetic wave radiation, vibration, electrostatics, electrodynamics, fluid flow, and elasticity, just to name a few. For this reason, the theoretical and numerical development of PDEs has been considered the core of applied mathematics, at least in the academic environment.

On the other hand, over the past two decades, we have witnessed a rapidly increasing volume of "information" contents to be processed and understood. With the recent advances of various high-tech fields and the popularity of social networking, the trend of exponential growth of easily accessible information is certainly going to continue well into the twenty-first century, and the bottleneck created by this information explosion will definitely require innovative solutions

from the scientific and engineering communities, particularly those technologists with better understanding of, and strong background in, applied mathematics. Today, "big data" is among the most pressing research directions. "Big data" research and development initiatives have been created by practically all Federal departments and agencies of the United States. It is noted, in particular, that on May 29, 2012, the Obama administration unveiled a "big data" initiative, announcing $200 million new R&D investments to help solve some of the nation's most pressing challenges, by improving our ability to extract knowledge and insights from large and complex collections of digital data. To launch the initiative, six Federal departments and agencies later announced more than $200 million in new commitments that promised to greatly improve the tools and techniques needed to access, organize, and glean discoveries from huge volumes of digital data. "Mathematics of big data" is therefore expected to provide innovative theory, methods, and algorithms to virtually every discipline, far beyond sciences and engineering, for processing, transmitting, receiving, understanding, and visualizing datasets, which could be very large or live in some high-dimensional spaces.

Of course the basic mathematical tools, particularly PDE models and methods, are always among the core of the mathematical tool-box of applied mathematics. But other theory and methods have been integrated in this tool-box as well. One of the most essential ideas is the notion of "frequency" of the data information. A contemporary of Gauss, by the name of Joseph Fourier, instilled this important concept to our study of physical phenomena by his innovation of trigonometric series representations, along with powerful mathematical theory and methods, which significantly expanded the core of the tool-box of applied mathematics. The frequency content of a given dataset facilitates the processing and understanding of the data information. Another important idea is the "multi-scale" structure of datasets. Less than three decades ago, with the birth of another exciting mathematical subject, called "wavelets", the dataset of information can be put in the wavelet domain for multi-scale processing as well. About half of this book is devoted to the study of the theories, methods, algorithms, and computational schemes of Fourier and wavelet analyses. In addition, other mathematical topics, which are essential to information processing but not commonly taught in a regular applied mathematics course, are discussed in this book. These include information coding, data dimensionality reduction, and data compression.

The objective of this textbook is to introduce the basic theory and methods in the tool-box of the core of applied mathematics, with a central scheme that addresses information processing with emphasis on manipulation of digital image data. Linear algebra is presented as linear analysis, with emphasis on spectral representation and principal component analysis (PCA), and with applications to data estimation and data dimensionality reduction. For data compression, the notion of entropy is introduced to quantify coding efficiency as governed by Shannon's Noiseless Coding theorem. Discrete Fourier transform (DFT), followed by the discussion of an efficient computational algorithm, called fast Fourier transform (FFT), as well as a real-valued version of the DFT, called discrete cosine

transform (DCT), are studied, with application to extracting frequency content of the given discrete dataset that facilitates reduction of the entropy and thus significant improvement of the coding efficiency. While DFT is obtained from discretization of the Fourier coefficient integral, four versions of discretization of the Fourier cosine coefficients yield the DCT-I, DCT-II, DCT-III and DCT-IV that are commonly used in applications. The integral version of DCT and Fourier coefficient sequences is called the Fourier transform (FT). Analogous to the Fourier series, the formulation of the inverse Fourier transform (IFT) is derived by applying the Gaussian function as sliding time-window for simultaneous time-frequency localization, with optimality guaranteed by the Uncertainty Principle. Both the Fourier series and Fourier transform are applied to solving certain PDEs.

In addition, local time-frequency basis functions are introduced in this textbook by discretization of the frequency-modulated sliding time-window function at the integer lattice points. Replacing the frequency modulation by modulation with the cosines avoids the Balian-Low stability restriction on the local time-frequency basis functions, with application to elimination of blocky artifacts caused by quantization of tiled DCT in image compression. In Chap. 8, multi-scale data analysis is introduced and compared with the Fourier frequency approach; the architecture of multiresolution approximation and analysis (MRA) is applied to the construction of wavelets and formulation of the multi-scale wavelet decomposition and reconstruction algorithms; and the lifting scheme is also introduced to reduce the computational complexity of these algorithms as well as implementation of filter banks. The final two chapters of this book are devoted to an in-depth study of wavelet analysis, including construction of bi-orthogonal wavelets, their regularities, and convergence of the cascade algorithms. The last three chapters alone, namely Chaps. 8–10, constitute a suitable textbook for a one-semester course in "Wavelet Analysis and Applications". For more details, a teaching guide is provided on pages xvi–xix.

Most chapters of this Applied Mathematics textbook have been tested in classroom teaching at both the undergraduate and graduate levels, and the final revision of the book manuscript was influenced by student feedback. The authors are therefore thankful to have the opportunity to teach from the earlier drafts of this book in the university environment, particularly at the University of Missouri-St. Louis. In writing this textbook, the authors have benefited from assistance of several individuals. In particular, they are most grateful to Margaret Chui for typing the earlier versions of the first six chapters, to Tom Li for drawing many of the diagrams and improving the artworks, and to Maryke van der Walt for proofreading the entire book and suggesting several changes. In addition, the first author would like to express his appreciation to the publisher, Atlantis Press, for their enthusiasm in publishing this book, and is grateful to Keith Jones and Zeger Karssen, in particular, for their unconditional trust and lasting friendship over the years. He would also like to take this opportunity to acknowledge the generous support from the U.S. Army Research Office and the National Geospatial-Intelligence Agency of

his research in the broad mathematical subject of complex and high-dimensional data processing. The second author is indebted to his family for their sacrifice, patience, and understanding, during the long hours in preparing and writing this book.

Charles K. Chui
Menlo Park, California

Qingtang Jiang
St. Louis Missouri

Contents

Teaching Guide

This book is an elementary and yet comprehensive textbook that covers most of the standard topics in Applied Mathematics, with a unified theme of applications to data analysis, data manipulation, and data compression. It is a suitable textbook, not only for most undergraduate and graduate Applied Mathematics courses, but also for a variety of Special Topics courses or seminars. The objective of this guide is to suggest several samples of such courses.

(1) A general "Applied Mathematics" course: Linear Spaces, Linear Analysis, Fourier Series, Fourier Transform, Partial Differential Equations, and Applications
(2) Applied Mathematics: with emphasis on computations and data compression
(3) Applied Mathematics: with emphasis on data analysis and representation
(4) Applied Mathematics: with emphasis on time-frequency analysis
(5) Applied Mathematics: with emphasis on time-frequency and multi-scale methods
(6) Applied Linear Algebra
(7) Applied Fourier Analysis
(8) Applied and Computational Wavelet Analysis
(9) Fourier and Wavelet Analyses

1. Teaching Guide (for a general Applied Mathematics Course)

```
                    ┌─────────────────┐
                    │   Chapter 1     │
                    └────────┬────────┘
                             ▼
                    ┌─────────────────┐
                    │   Chapter 2     │
                    └────────┬────────┘
                             ▼
                    ┌─────────────────┐
                    │   Chapter 3     │
                    └────────┬────────┘
                             ▼
                    ┌─────────────────┐
                    │   Chapter 6     │◄──────────────┐
                    └────────┬────────┘               │
                             ▼                        │
              ┌──────────────────────────┐            │
              │  Chapter 7: 7.1, 7.2, 7.3│            │
              └──────────────────────────┘            │
      ┌────────────────────────────────────────┐      │
      │         Chapters 4: 4.1, 4.2           │──────┘
      │  Supplementary materials on computations│
      └────────────────────────────────────────┘
```

2. Teaching Guide (Emphasis on computations and data compressions)

```
                    ┌─────────────────┐
                    │   Chapter 2     │◄──────────────┐
                    └────────┬────────┘               │
                             ▼                        │
                    ┌─────────────────┐               │
                    │   Chapter 3     │               │
                    └────────┬────────┘               │
                             ▼                        │
                    ┌─────────────────┐               │
                    │   Chapter 4     │               │
                    └────────┬────────┘               │
                             ▼                        │
                    ┌─────────────────┐               │
                    │   Chapter 5     │               │
                    └────────┬────────┘               │
                             ▼                        │
                    ┌─────────────────┐               │
                    │   Chapter 6     │               │
                    └─────────────────┘               │
      ┌────────────────────────────────────────┐      │
      │  Chapter 1: Background material, if needed│───┘
      └────────────────────────────────────────┘
```

3. Teaching Guide (Emphasis on data analysis and data representations)

```
            ┌──────────────────┐
            │    Chapter 2     │◄─────────────┐
            └──────────────────┘              │
                     │                        │
                     ▼                        │
            ┌──────────────────┐              │
            │    Chapter 3     │              │
            └──────────────────┘              │
                     │                        │
                     ▼                        │
            ┌──────────────────┐              │
            │    Chapter 4     │              │
            └──────────────────┘              │
                     │                        │
                     ▼                        │
            ┌──────────────────┐              │
            │    Chapter 6     │              │
            └──────────────────┘              │
                     │                        │
                     ▼                        │
       ┌─────────────────────────────┐        │
       │  Chapter 7: 7.1, 7.2, 7.3   │        │
       └─────────────────────────────┘        │
                                              │
   ┌──────────────────────────────────────────┤
   │  Chapter 1: Background material, if needed│
   └───────────────────────────────────────────┘
```

4. Teaching Guide (Emphasis on time-frequency analysis)

```
            ┌──────────────────┐
            │    Chapter 6     │◄─────────────┐
            └──────────────────┘              │
                     │                        │
                     ▼                        │
       ┌─────────────────────────────────┐    │
       │  Chapter 7: 7.1, 7.2, 7.4, 7.5  │    │
       └─────────────────────────────────┘    │
                     │                        │
                     ▼                        │
         ┌──────────────────────────┐          │
         │    Chapter 8: 8.1 – 8.2  │          │
         └──────────────────────────┘          │
                                              │
   ┌──────────────────────────────────────────┤
   │ Chapter 4: 4.1 on background material, if needed│
   └───────────────────────────────────────────┘
```

5. Teaching Guide (Emphasis on time-frequency/time-scale methods)

```
┌─────────────────────────────┐
│   Chapter 4: 4.1 – 4.2      │
└─────────────────────────────┘
              │
              ▼
┌─────────────────────────────┐
│   Chapter 6: 6.1 – 6.2      │
└─────────────────────────────┘
              │
              ▼
┌───────────────────────────────────┐
│   Chapter 7: 7.1, 7.2, 7.4, 7.5   │
└───────────────────────────────────┘
              │
              ▼
      ┌─────────────────┐
      │   Chapter 8     │
      └─────────────────┘
              │
              ▼
┌─────────────────────────────┐
│   Chapter 9: 9.5 only       │
└─────────────────────────────┘
```

6. Teaching Guide (Applied Linear Algebra)

```
┌─────────────────────────────┐
│   Chapter 1: 1.2 – 1.5      │
└─────────────────────────────┘
              │
              ▼
      ┌─────────────────┐
      │   Chapter 2     │
      └─────────────────┘
              │
              ▼
      ┌─────────────────┐
      │   Chapter 3     │
      └─────────────────┘
              │
              ▼
      ┌─────────────────┐
      │   Chapter 4     │
      └─────────────────┘
              │
              ▼
      ┌─────────────────┐
      │   Chapter 5     │
      └─────────────────┘
```

7. Teaching Guide (Applied Fourier Analysis)

```
┌─────────────────────────────┐
│   Chapter 6: 6.1 – 6.4      │
└─────────────────────────────┘
              │
              ▼
┌─────────────────────────────┐
│   Chapter 4: 4.1 – 4.2      │
└─────────────────────────────┘
              │
              ▼
      ┌─────────────────┐
      │   Chapter 5     │
      └─────────────────┘
              │
              ▼
┌───────────────────────────────────┐
│   Chapter 7: 7.1, 7.2, 7.4, 7.5   │
└───────────────────────────────────┘
```

8. Teaching Guide (Applied and Computational Wavelet Analysis)

```
            ┌─────────────────────────────┐
            │   Chapter 7: 7.1, 7.2, 7.4  │
            └─────────────────────────────┘
                          │
                          ▼
               ┌────────────────────┐
               │     Chapter 8      │
               └────────────────────┘
                          │
                          ▼
               ┌────────────────────┐
               │     Chapter 9      │
               └────────────────────┘
                          │
                          ▼
              ┌──────────────────────┐
              │     Chapter 10       │
              └──────────────────────┘
```

9. Teaching Guide (Fourier and Wavelet Analyses)

```
            ┌─────────────────────────────┐
            │     Chapter 6: 6.1 – 6.4    │
            └─────────────────────────────┘
                          │
                          ▼
        ┌──────────────────────────────────────┐
        │    Chapter 7: 7.1, 7.2, 7.4, 7.5     │
        └──────────────────────────────────────┘
                          │
                          ▼
               ┌────────────────────┐
               │     Chapter 8      │
               └────────────────────┘
                          │
                          ▼
               ┌────────────────────┐
               │     Chapter 9      │
               └────────────────────┘
                          │
                          ▼
               ┌────────────────────┐
               │     Chapter 10     │
               └────────────────────┘
```

Figures

Chapter 1
Linear Spaces

The two most basic operations on a given set $\mathbb{V}$ of functions are addition of any two functions in $\mathbb{V}$ and multiplication of any function in $\mathbb{V}$ by a constant. The concept of "vector spaces" to be discussed in some details in the first section of this chapter is to specify the so-called closure properties of the set $\mathbb{V}$, in that these two operations never create any function outside the set $\mathbb{V}$. More precisely, the set of constants that are used for multiplication to the functions in $\mathbb{V}$ must be a "scalar field" denoted by $\mathbb{F}$, and $\mathbb{V}$ will be called a vector space over the field $\mathbb{F}$. The only (scalar) fields considered in this book are the set $\mathbb{R}$ of real numbers as well as its subset $\mathbb{Q}$ of rational numbers and its superset $\mathbb{C}$ of complex numbers. Typical examples of vector spaces over $\mathbb{F}$ include the vector spaces $\mathbb{R}^{m,n}$ and $\mathbb{C}^{m,n}$ of $m \times n$ matrices of real numbers and of complex numbers, over the scalar fields $\mathbb{F} = \mathbb{R}$ and $\mathbb{F} = \mathbb{C}$, respectively, where m and n are integers with $m, n \geq 1$. If $m = 1$, then the vector space of matrices becomes the familiar space of row vectors, while for $n = 1$ we have the familiar space of column vectors.

The vector space of infinite sequences and that of piecewise continuous functions are studied in the second section, Sect. 1.2. The term "vector" will be used throughout the entire book for any member of the vector space under consideration. Hence, a vector could be a matrix, an infinite sequence, a function, and so forth. For any p with $0 \leq p \leq \infty$, in order to show that the sequence subspace ℓ_p and function subspace $\widetilde{L}_p(J)$ are closed under addition, we will establish Minkowski's inequalities, also called triangle inequality in the general setting. For this purpose, the inequalities of Hölder are first derived by applying Young's inequality, which is also discussed in this section. For this preliminary discussion, the functions in $\widetilde{L}_p(J)$ are restricted to piecewise continuous functions on an interval J which is allowed to be either bounded or unbounded. Extension to "measurable functions" is delayed to the last section, Sect. 1.5, of the chapter.

The third section, Sect. 1.3, of this chapter is devoted to the study of an important measurement tool for vector spaces, namely: the inner product. Since the inner product of any two vectors is a constant in the scalar field, it is also called "scalar product", and particularly the "dot product" for the Euclidean spaces $\mathbb{R}^m = \mathbb{R}^{m,1}$ and $\mathbb{C}^m = \mathbb{C}^{m,1}$. A vector space $\mathbb{V}$ endowed with some inner-product measurement is called an inner-product space. The notion of "norm" of a vector in an inner-product space is introduced to facilitate our discussion of the most important property of the inner-product measurement, namely: the Cauchy-Schwarz inequality, which is the generalization of Hölder's inequality for ℓ_2 and $\widetilde{L}_2(J)$, derived in Sect. 1.2, to the general abstract setting. When the norm is used to measure the length of a vector in a real inner-product space $\mathbb{V}$ over the scalar field $\mathbb{R}$, the Cauchy-Schwarz inequality can be applied to define the angle between any two vectors in $\mathbb{V}$. This concept motivates the definition of orthogonality, which obviously extends to all inner-product spaces $\mathbb{V}$, including those over the scalar field $\mathbb{C}$. As an application, the notion of "orthogonal complement" of any subspace $\mathbb{W} \in \mathbb{V}$ is introduced for the study of applications in later chapters of this book.

The concepts of linear independence and bases from an elementary course in Linear Algebra are reviewed and extended in Sect. 1.4 to infinite-dimensional vector spaces, such as the vector space ℓ_p of infinite sequences and the function space $\widetilde{L}_p(J)$, for any p with $0 \le p \le \infty$, studied in Sect. 1.2. Another objective of Sect. 1.4 is to apply the inner product of an inner-product space $\mathbb{V}$ discussed in Sect. 1.3 to introduce the notion of orthonormal basis of $\mathbb{V}$ and that of orthogonal projection from $\mathbb{V}$ to any of its subspaces $\mathbb{W}$, as well as best approximation of vectors in $\mathbb{V}$ from $\mathbb{W}$. In addition, we will give an in-depth discussion of the so-called Gram-Schmidt process for constructing an orthonormal set of vectors from any linearly independent collection, while preserving the algebraic span of the linearly independent set.

The notion of the norm measurement defined by an inner product in Sect. 1.3 is extended to the general abstract setting in Sect. 1.5. A vector space endowed with a norm is called a normed space (or normed linear space). Important examples of normed spaces include ℓ_p and $\widetilde{L}_p(J)$ for any p, with $1 \le p \le \infty$, introduced in Sect. 1.2. While the length of a vector as defined by its norm can also be used to measure the distance between any two vectors in a normed space, the notion of "metric" measurement from an elementary course in Set Theory for measuring the distance between two points in a point-set can be applied without the vector space structure. Important examples of metric spaces that are not normed spaces are ℓ_p and $\widetilde{L}_p(J)$ for any p, with $0 \le p < 1$. With the metric, the definition of Cauchy sequences can be used to introduce the concept of completeness. For example, the completion of $\widetilde{L}_p(J)$ is $L_p(J)$, for any p with $0 \le p \le \infty$, by extending the collection of piecewise continuous functions to the (Lebesgue) measurable functions. A complete normed space is called a Banach space and a complete inner-product space iss called a Hilbert space.

1.1 Vector Spaces

For the definition of a vector space, we need the notion of its companion scalar field. Throughout this book, the set of all real numbers is denoted by $\mathbb{R}$. Under the operations of addition/subtraction, multiplication, and division by non-zero real numbers, the set $\mathbb{R}$ is a field. For the purpose of our discussion of vector spaces, any field will be called a scalar field, and any element of the field will be called a scalar.

In this book, the set of natural numbers is denoted by $\mathbb{N}$. Both $\mathbb{N}$ and the notation $\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$ for the set of all integers will be used frequently in our discussions. For example, the set of all fractions can be represented by

$$\mathbb{Q} = \left\{ \frac{p}{q} : p \in \mathbb{Z}, q \in \mathbb{N} \right\}.$$

A fraction is also called a rational number, and the set $\mathbb{Q}$ of rational numbers is also a scalar field, and hence a sub-field of $\mathbb{R}$, since addition/subtraction and multiplication of two fractions, as well as division of a fraction by any non-zero fraction remain to be a fraction. Hence, the set $\mathbb{Q}$ can also be used as a scalar field in the discussion of vector spaces.

To better understand the set $\mathbb{Q}$ of rational numbers, let us first recall that a real number with terminating decimal representation is in $\mathbb{Q}$. Indeed, as an example, the real number 2.31 can be written as $2 + \frac{31}{100} = \frac{231}{100}$, which is a fraction. A less obvious fraction representation of certain real numbers is one that has infinitely many digits in its decimal representation, but after writing out finitely many digits, the decimal representation repeats indefinitely the same pattern of some finite sequence. This pattern therefore constitutes a period of the decimal representation. To illustrate this statement, consider the real number $b = 2.3121212\ldots$, with the periodic pattern of 12 tacked on to the decimal representation 2.3. To see that b can be expressed as a fraction, let us first consider another real number

$$x = 0.1212 \cdots .$$

Since the pattern of the two digits 12 in x repeats indefinitely, we may multiply x by $10^2 = 100$ to obtain the equation

$$100x = 12.121212 \cdots = 12 + x$$

with unique solution given by

$$x = \frac{12}{100 - 1} = \frac{12}{99},$$

which is a fraction. Returning to b, we can compute its fraction representation as follows:

$$b = 2.3 + 0.0121212 \cdots = 2.3 + \frac{x}{10}$$

$$= 2 + \frac{3}{10} + \frac{12}{990} = \frac{2289}{990}.$$

More generally, it can be shown that if

$$x = 0.c_1 c_2 \cdots c_s c_1 c_2 \cdots c_s c_1 c_2 \cdots c_s \cdots$$

(and the notation $x = 0.\overline{c_1 c_2 \cdots c_s}$ is used to indicate that the pattern $c_1 c_2 \cdots c_s$ is repeated indefinitely), then x may be written as the fraction:

$$0.\overline{c_1 c_2 \cdots c_s} = \frac{c_1 c_2 \cdots c_s}{10^s - 1}$$

(see Exercise 3). Thus, if a real number y has the decimal representation

$$y = a_1 \cdots a_n . b_1 \cdots b_m c_1 c_2 \cdots c_s c_1 c_2 \cdots c_s c_1 c_2 \cdots c_s \cdots ,$$

then y can be written as a fraction by adding three terms as follows:

$$a_1 \cdots a_n + \frac{b_1 \cdots b_m}{10^m} + \frac{c_1 c_2 \cdots c_s}{(10^s - 1)10^m}.$$

On the other hand, the decimal representation for most real numbers does not have a periodic pattern, and hence "almost all" real numbers do not have fraction representations and therefore, are not rational numbers. A real number which is not a rational is called an irrational number. For example, the numbers e (in honor of Euler, and used to define the natural logarithm), π (the ratio of the circumference of a circle with its diameter, introduced by Archimedes of Syracuse), and $\sqrt{2}$ are real numbers that are not in $\mathbb{Q}$. In fact, it is not difficult to show that if $a > 1$ is not a perfect square (that is, $a \neq b^2$ for any integer $b > 1$), then $\sqrt{a}$ is an irrational number (or equivalently, different from a fraction). These numbers do not have periodic patterns in their decimal representations. For example, the first few decimals of e, π, and $\sqrt{2}$ are:

$$e = 2.718281828459045 \cdots ,$$
$$\pi = 3.141592653589793 \cdots ,$$
$$\sqrt{2} = 1.414213652373095 \cdots .$$

Since the decimal representations of "almost all" real numbers are so irregular, it seems to be impossible to enumerate or "count" the set of all real numbers. In the mathematical language, a set is said to be **countable**, if either it has a finite number of members (also called elements) or it can be counted by assigning each member of the set to one and only one natural number, while a set is said to be uncountable

if the set is not countable. More precisely, a set S is countable if and only if it can be listed by assigning some natural number to each of its members as follows:

$$S = \{s_1, s_2, s_3, \ldots, s_n\} \text{ or } S = \{s_1, s_2, s_3, \ldots\},$$

for some $n \in \mathbb{N}$.

$\boxed{\text{The set } \mathbb{Q} \text{ of rational numbers is countable}}$ To show this, we first point out that the union of two countable sets is also countable (see Exercise 4), so that it is sufficient to prove that the set of positive rational numbers is countable. Allowing redundancy, we display the set of all positive fractions p/q in a triangular array in the figure below, where p runs over all natural numbers $1, 2, 3, \ldots$, from northwest to southeast, along each line of the array with slope -1, and q runs over all natural numbers $1, 2, 3, \ldots$ from the top along each vertical line of the array.

$$
\begin{array}{l}
1/1 \\
1/2 \ 2/1 \\
1/3 \ 2/2 \ 3/1 \\
1/4 \ 2/3 \ 3/2 \ 4/1 \\
1/5 \ 2/4 \ 3/3 \ 4/2 \ 5/1 \\
1/6 \ 2/5 \ 3/4 \ 4/3 \ 5/2 \ 6/1 \\
\cdots \ \cdots \ \cdots \ \cdots \ \cdots \ \cdots \ \cdots
\end{array}
$$

Then by deleting redundant representation of the same fractions, all positive rational numbers are listed in a unique manner on this triangular array (counting row by row from the top) as follows:

$$\frac{1}{1}, \frac{1}{2}, \frac{2}{1}, \frac{1}{3}, \frac{3}{1}, \frac{1}{4}, \frac{2}{3}, \frac{3}{2}, \frac{4}{1}, \frac{1}{5}, \frac{5}{1}, \frac{1}{6}, \frac{2}{5}, \ldots,$$

where redundancy is avoided by skipping $2/2, 2/4, \ldots, 3/3, 3/6, \ldots, 4/4, 4/6, \ldots$, and so forth.

Remark 1 $\boxed{\text{The set } \mathbb{Q} \text{ of rational numbers has measure } = 0}$ This statement means that the set $\mathbb{Q}$ is contained in the union of some intervals with arbitrarily small total length.

To show this, let $\mathbb{Q} = \{r_1, r_2, \ldots\}$ be the set of all rational numbers, listed in any desirable order, and place all the rational numbers on the real line (or x-axis). Now let ϵ be an arbitrarily small positive number. Then for each $k = 1, 2, \ldots$, consider the interval $I_k = [r_k - \epsilon \, 2^{-(k+1)}, r_k + \epsilon \, 2^{-(k+1)}]$ with center at r_k. Hence, the set $\mathbb{Q}$ of all rational numbers is contained in the union of the intervals $I_1, I_2, \ldots$. Therefore, since the total length of these intervals is equal to

$$\frac{\epsilon}{2} + \frac{\epsilon}{2^2} + \frac{\epsilon}{2^3} + \cdots = \epsilon,$$

the set $\mathbb{Q}$ of all rational numbers is contained in a set of "measure" less than the arbitrarily small positive number ϵ. Such sets are said to have measure $= 0$.

Observe that since the real line has infinite length, the set $\mathbb{R}\backslash\mathbb{Q}$ of irrational numbers must be uncountable, and so is the superset $\mathbb{R}$ of $\mathbb{R}\backslash\mathbb{Q}$. So, in comparison with $\mathbb{Q}$, the set of irrational numbers is much larger.   ∎

Remark 2 | **The notions of "almost all" and "almost everywhere"** | In view of the fact that the set of rational numbers has measure zero, we say that "almost all" real numbers are irrational numbers. Extending this concept to functions defined on an interval J, we say that two functions f and g are equal almost everywhere on J, if $f(x) = g(x)$ for all $x \in J\backslash S$, for some subset S of the interval J with measure zero, and we write $f = g$ a.e. Of course S could be the empty set.   ∎

Remark 3 Real numbers that are not fractions must be quantized (for example, by truncation followed by rounding off the last digit) for fixed point computation to avoid using symbolic implementation. Unfortunately, the quantization process is irreversible. For example, to multiply $\sqrt{2}$ by 3 without symbolic calculation, $\sqrt{2}$ can be quantized to 1.4142 before multiplying by 3, yielding:

$$\sqrt{2} \times 3 \approx 1.4142 \times 3 = 4.2426.$$

But when this number is later squared, we will only obtain $4.2426^2 = 17.99965476$, which is different from the exact value $(\sqrt{2} \times 3)^2 = 2 \times 9 = 18$. On the other hand, the quantization process is a key step in such data compression applications that require high compression ratio (or low bit-rate). This topic will be discussed in Chap. 5.   ∎

The scalar field $\mathbb{R}$ of real numbers can be extended to the field $\mathbb{C}$ of complex numbers, so that $\mathbb{C}$ is a super-field of $\mathbb{R}$, or equivalently, $\mathbb{R}$ is a sub-field of $\mathbb{C}$. For this purpose, recall the imaginary number i defined by $i^2 = -1$; that is, i and $-i$ are the two roots of the equation $x^2 + 1 = 0$. Then a complex number $c \in \mathbb{C}$ is defined by

$$c = a + ib, \quad a, b \in \mathbb{R}.$$

Hence, while a real number $a \in \mathbb{R}$ is customarily indicated by a "point" on the "real line", the plot of a complex number c is a point on a plane, called the complex plane, also denoted by $\mathbb{C}$, with the horizontal axis to be called the real axis and the vertical axis to be called the imaginary axis.

For the complex number $c = a + ib$, the real number a is called the real part of c, and the other real number b is called the imaginary part of c, denoted, respectively, by

$$a = \operatorname{Re} c, \quad b = \operatorname{Im} c.$$

The magnitude $|c|$ of c is defined by

$$|c| = \sqrt{a^2 + b^2}.$$

If $c \neq 0$, it can be written in the polar form

$$c = \rho e^{i\theta}$$

for $\rho = |c|$ and θ defined by

$$\cos\theta = a/\rho, \ \sin\theta = b/\rho,$$

with the value of θ restricted to the interval $[0, 2\pi)$. Observe that without this restriction of θ, the above definition yields Euler's identity:

$$e^{i\theta} = \cos\theta + i\sin\theta,$$

which is valid for all real numbers θ.

For $c = a + ib$, the complex number

$$\bar{c} = a - ib,$$

with b in c replaced by $-b$, is called the complex conjugate of c. See Fig.1.1.

By applying the definition $i^2 = -1$, we may conclude that

$$|c|^2 = c\,\bar{c} = a^2 + b^2;$$
$$\frac{1}{c} = \frac{\bar{c}}{|c|^2} = \frac{a - ib}{a^2 + b^2} = \frac{a}{a^2 + b^2} - i\frac{b}{a^2 + b^2}.$$

While the "real line", also denoted by $\mathbb{R}$, is the set of real numbers that constitute all the "points" that lie on some horizontal line, its extension to the "plane"

$$\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$$

is the set of ordered pairs $\mathbf{x} = (x_1, x_2)$, where $x_1, x_2 \in \mathbb{R}$. For example, $\mathbf{x} = (1.5, -3)$ is a "point" in the plane $\mathbb{R}^2$ with first component (also called coordinate) being 1.5 and the second component being -3. Observe that when the two components are interchanged, we have a different point on the plane, namely:
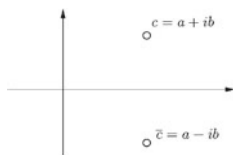

Fig. 1.1 Plots of c and $\bar{c}$ in the complex plane

$$(1.5, -3) \neq (-3, 1.5).$$

Similarly, the plane $\mathbb{R}^2$ is extended to the three-dimensional space $\mathbb{R}^3$ by

$$\mathbf{x} = (x_1, x_2, x_3) \in \mathbb{R}^3,$$

where $x_1, x_2, x_3 \in \mathbb{R}$. In general, for each integer $n \geq 2$, the n-dimensional space $\mathbb{R}^n$ is defined by

$$\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n, \tag{1.1.1}$$

where $x_1, \ldots, x_n \in \mathbb{R}$. Analogously, the n-dimensional space of "vectors" with complex-valued components is denoted by $\mathbb{C}^n$, with

$$\mathbf{z} = (z_1, \ldots, z_n) \in \mathbb{C}^n, \tag{1.1.2}$$

where $z_1, \ldots, z_n \in \mathbb{C}$.

In the above discussion, we have used the words "vector" and "space" loosely; with the intention of motivating the introduction of the concept of "vector spaces", as follows.

Definition 1 | **Vector space** | *Let $\mathbb{F}$ be a scalar field such as $\mathbb{C}$, $\mathbb{R}$ or $\mathbb{Q}$. A nonempty collection $\mathbb{V}$ of elements (called "vectors") together with two operations (one called "addition" and the other "scalar multiplication"), is said to be a vector space over the scalar field $\mathbb{F}$, if the following operations are satisfied.*

(a) *Closure under vector addition:*

$$\mathbf{x}, \mathbf{y} \in \mathbb{V} \Rightarrow \mathbf{x} + \mathbf{y} \in \mathbb{V}.$$

(b) *Closure under scalar multiplication:*

$$\mathbf{x} \in \mathbb{V} \text{ and } a \in \mathbb{F} \Rightarrow a\mathbf{x} \in \mathbb{V}.$$

(c) *Rules for vector addition and scalar multiplication:*

   (i) $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$, *for all* $\mathbf{x}, \mathbf{y} \in \mathbb{V}$,
   (ii) $(\mathbf{x} + \mathbf{y}) + \mathbf{z} = \mathbf{x} + (\mathbf{y} + \mathbf{z})$, *for all* $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{V}$,
   (iii) *there is a zero element* $\mathbf{0}$ *(called zero vector) in* $\mathbb{V}$ *such that* $\mathbf{x} + \mathbf{0} = \mathbf{x}$, *for all* $\mathbf{x} \in \mathbb{V}$,
   (iv) *for each* $\mathbf{x} \in \mathbb{V}$, $-\mathbf{x} = (-1)\mathbf{x}$ *satisfies* $\mathbf{x} + (-\mathbf{x}) = \mathbf{0}$ *and* $a(\mathbf{x} + \mathbf{y}) = (\mathbf{x} + \mathbf{y})a = a\mathbf{x} + a\mathbf{y}$ *for all* $a \in \mathbb{F}$ *and* $\mathbf{x}, \mathbf{y} \in \mathbb{V}$,
   (v) $(a + b)\mathbf{x} = a\mathbf{x} + b\mathbf{x}$, *for all* $a, b \in \mathbb{F}$ *and* $\mathbf{x} \in \mathbb{V}$,
   (vi) $a(b\mathbf{x}) = (ab)\mathbf{x}$, *for all* $a, b \in \mathbb{F}$ *and* $\mathbf{x} \in \mathbb{V}$, *and*
   (vii) $1\mathbf{x} = \mathbf{x}$ *for all* $\mathbf{x} \in \mathbb{V}$.

Remark 4 The closure properties (a) and (b) of a vector space $\mathbb{V}$ over a scalar field $\mathbb{F}$ can be unified as a single condition, namely :

$$a, b \in \mathbb{F} \ \text{ and } \ \mathbf{x}, \mathbf{y} \in \mathbb{V} \Rightarrow a\mathbf{x} + b\mathbf{y} \in \mathbb{V}. \tag{1.1.3}$$

Indeed, by setting $a = b = 1$ in (1.1.3), we have (a) in Definition 1; and by setting $b = 0$, (b) is satisfied. Conversely, for $a, b \in \mathbb{F}, \mathbf{x}, \mathbf{y} \in \mathbb{V}$, it follows from (b) that $a\mathbf{x}, b\mathbf{y} \in \mathbb{V}$. This, together with (a), implies that $a\mathbf{x} + b\mathbf{y} \in \mathbb{V}$, as desired.   ∎

From (iv) in (c) in Definition 1, we have $\mathbf{y} - \mathbf{x} = \mathbf{y} + (-\mathbf{x})$ is in $\mathbb{V}$ and this operation is called the "difference" of $\mathbf{y}$ and $\mathbf{x}$.

The following result should be easy to verify.

Theorem 1 $\boxed{\textbf{Vector spaces } \mathbb{R}^n \textbf{ and } \mathbb{C}^n}$ *Let $\mathbb{C}^n$, where $n \geq 1$ is a positive integer, be defined as in* (1.1.2), *with vector addition and scalar multiplication defined by*

$$\begin{aligned} \mathbf{x} + \mathbf{y} &= (x_1, \ldots, x_n) + (y_1, \ldots, y_n) \\ &= (x_1 + y_1, \ldots, x_n + y_n) \end{aligned}$$

and

$$a\mathbf{x} = a(x_1, \ldots, x_n) = (ax_1, \ldots, ax_n).$$

Then $\mathbb{C}^n$ is a vector space over the scalar field $\mathbb{C}$. The same statement is valid if $\mathbb{C}$ and $\mathbb{C}^n$ are replaced by $\mathbb{R}$ and $\mathbb{R}^n$, respectively. The vector spaces $\mathbb{R}^n$ and $\mathbb{C}^n$ are called Euclidean spaces.

Next, let m and n be integers with $m, n \geq 1$. Then the definition of the vector space $\mathbb{C}^n$ in (1.1.2) can be extended to the set $\mathbb{C}^{m,n}$ of $m \times n$ matrices, defined by

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \in \mathbb{C}^{m,n}, \tag{1.1.4}$$

where $a_{jk} = a_{j,k}, 1 \leq j \leq m$ and $1 \leq k \leq n$, are complex numbers (observe that the comma that separates j and k is often omitted for convenience). Similarly, $\mathbb{R}^{m,n}$ denotes the set of $m \times n$ real matrices (matrices whose entries consist entirely of real numbers). By extending the operations of vector addition and scalar multiplication for $\mathbb{C}^n$ in Theorem1 to $\mathbb{C}^{m,n}$, namely:

$$
\begin{aligned}
A + B \;&=\;
\begin{bmatrix}
a_{11} & a_{12} & \cdots & a_{1n} \\
a_{21} & a_{22} & \cdots & a_{2n} \\
\vdots & \vdots & \cdots & \vdots \\
a_{m1} & a_{m2} & \cdots & a_{mn}
\end{bmatrix}
+
\begin{bmatrix}
b_{11} & b_{12} & \cdots & b_{1n} \\
b_{21} & b_{22} & \cdots & b_{2n} \\
\vdots & \vdots & \cdots & \vdots \\
b_{m1} & b_{m2} & \cdots & b_{mn}
\end{bmatrix} \\[2mm]
&=\;
\begin{bmatrix}
a_{11}+b_{11} & a_{12}+b_{12} & \cdots & a_{1n}+b_{1n} \\
a_{21}+b_{21} & a_{22}+b_{22} & \cdots & a_{2n}+b_{2n} \\
\vdots & & \cdots & \vdots \\
a_{m1}+b_{m1} & a_{m2}+b_{m2} & \cdots & a_{mn}+b_{mn}
\end{bmatrix}
\end{aligned}
\tag{1.1.5}
$$

and

$$
aA = a
\begin{bmatrix}
a_{11} & \cdots & a_{1n} \\
\vdots & \cdots & \vdots \\
a_{m1} & \cdots & a_{mn}
\end{bmatrix}
=
\begin{bmatrix}
aa_{11} & \cdots & aa_{1n} \\
\vdots & \cdots & \vdots \\
aa_{m1} & \cdots & aa_{mn}
\end{bmatrix},
\tag{1.1.6}
$$

we have the following result.

Theorem 2 $\boxed{\text{Vector spaces } \mathbb{C}^{m,n} \text{ and } \mathbb{R}^{m,n}}$ *The set $\mathbb{C}^{m,n}$ of $m \times n$ matrices of complex numbers defined by (1.1.4) and operations as in (1.1.5) and (1.1.6) is a vector space over the scalar field $\mathbb{C}$. This statement remains valid if $\mathbb{C}^{m,n}$ and $\mathbb{C}$ are replaced by $\mathbb{R}^{m,n}$ and $\mathbb{R}$, respectively.*

Definition 2 $\boxed{\text{Subspaces}}$ *A nonempty subset $\mathbb{W}$ of a vector space $\mathbb{V}$ over $\mathbb{F}$ is called a subspace of $\mathbb{V}$, if $\mathbb{W}$ is also a vector space over $\mathbb{F}$ with the same addition and scalar multiplication operations of $\mathbb{V}$.*

Since $\mathbb{W} \subset \mathbb{V}$, the operations in (i)–(ii) and (iv)–(vii) of Definition 1 are already valid for vectors in $\mathbb{W}$. In addition, since $\mathbb{W}$ is nonempty, there exists $\mathbf{v}_0 \in \mathbb{W}$. Thus, from (v) in Definition 1, we see $\mathbf{0} = 0\mathbf{v}_0 \in \mathbb{W}$. That is, the zero vector $\mathbf{0}$ is always in $\mathbb{W}$. Therefore, the operation (iii) applies to $\mathbb{W}$. Hence, to find out if $\mathbb{W}$ is a subspace of $\mathbb{V}$, it is sufficient to verify if $\mathbb{W}$ satisfies the closure properties (a) and (b) in Definition 1, or equivalently, the condition (1.1.3).

Example 1 Let $\mathbb{W}_1 = \{(c, 0, d) : c, d \in \mathbb{R}\}$ and $\mathbb{W}_2 = \{(c, 1, d) : c, d \in \mathbb{R}\}$. Show that $\mathbb{W}_1$ is a subspace of $\mathbb{R}^3$ but $\mathbb{W}_2$ is not (although it is a subset of $\mathbb{R}^3$).

Solution Clearly, $\mathbb{W}_1$ is nonempty. As remarked above, it is enough to show that $\mathbb{W}_1$ satisfies (1.1.3).

For all $a, b \in \mathbb{R}$ and any $\mathbf{x} = (x_1, 0, x_3)$, $\mathbf{y} = (y_1, 0, y_3)$ in $\mathbb{W}_1$, we have

$$
\begin{aligned}
a\mathbf{x} + b\mathbf{y} &= a(x_1, 0, x_3) + b(y_1, 0, y_3) \\
&= (ax_1, 0, ax_3) + (by_1, 0, by_3) \\
&= (ax_1 + by_1, 0, ax_3 + by_3),
\end{aligned}
$$

which is in $\mathbb{W}_1$, by setting $c = ax_1 + by_1$ and $d = ax_3 + by_3$.

On the other hand, since $a(x_1, 1, x_3) = (ax_1, a, ax_3) \notin \mathbb{W}_2$ for any real number $a \neq 1$, it follows that $\mathbb{W}_2$ is not a subspace of $\mathbb{R}^3$. ∎

Let $A = [a_{jk}]_{1 \leq j,k \leq n} \in \mathbb{C}^{n,n}$ be a square matrix. Then A is called an **upper-triangular matrix** if $a_{jk} = 0$ for all $j > k$; and A is a called a **lower-triangular matrix** if $a_{jk} = 0$ for all $j < k$. The set of all $n \times n$ upper-triangular matrices is a subspace of $C^{n,n}$; and so is the set of all $n \times n$ lower-triangular matrices.

A square matrix $A_n = [a_{jk}]_{1 \leq j,k \leq n}$ is called a **diagonal matrix**, if all entries that are not on the main diagonal (with $j = k$) are equal to 0. The notation for a diagonal matrix A is

$$A = \text{diag}\{a_{11}, \ldots, a_{nn}\}.$$

Clearly, a diagonal matrix is an upper-triangular matrix and also a lower-triangular matrix.

Let $\mathbb{F}$ be the scalar field $\mathbb{Q}$, $\mathbb{R}$, or $\mathbb{C}$ and $P(t) = a_n t^n + a_{n-1} t^{n-1} + \cdots + a_1 t + a_0$ denote a polynomial in the variable t, where $a_j \in \mathbb{F}$ for $j = 0, \ldots, n$ and $a_n \neq 0$. Then $P(t)$ is called a polynomial of degree n with leading coefficient a_n, and the notation $\deg P = n$ is used for the degree of the polynomial. We also denote by Π the set of all polynomials. Then it is easy to show that Π is a vector space over $\mathbb{F}$ under the standard definitions of addition and scalar multiplication of functions. For each integer $n \geq 0$, let $\Pi_n = \{0\} \cup \{$all polynomials with degree $\leq n\}$. Then it is also easy to show that Π_n is a subspace of Π. ∎

We conclude this section by discussing the sum and intersection of two subspaces. Let $\mathbb{U}$ and $\mathbb{W}$ be two subspaces of a vector space $\mathbb{V}$. Define $\mathbb{U} + \mathbb{W}$, called the sum of $\mathbb{U}$ and $\mathbb{W}$, by

$$\mathbb{U} + \mathbb{W} = \{\mathbf{u} + \mathbf{w} : \mathbf{u} \in \mathbb{U}, \mathbf{w} \in \mathbb{W}\};$$

and the intersection of $\mathbb{U}$ and $\mathbb{W}$ by

$$\mathbb{U} \cap \mathbb{W} = \{\mathbf{v} : \mathbf{v} \in \mathbb{U} \text{ and } \mathbf{v} \in \mathbb{W}\}.$$

Then both $\mathbb{U} + \mathbb{W}$ and $\mathbb{U} \cap \mathbb{W}$ are subspaces of $\mathbb{V}$. Here, we give the proof for $\mathbb{U} + \mathbb{W}$, and leave the proof for $\mathbb{U} \cap \mathbb{W}$ as an exercise (see Exercise 14).

First, since $\mathbf{0} \in \mathbb{U}$ and $\mathbf{0} \in \mathbb{W}$, $\mathbf{0} = \mathbf{0} + \mathbf{0} \in \mathbb{U} + \mathbb{W}$. Thus $\mathbb{U} + \mathbb{W}$ is not empty. As remarked above, it suffices to prove that $\mathbb{U} + \mathbb{W}$ satisfies (1.1.3). Suppose $a, b \in \mathbb{F}$ and $\mathbf{x}, \mathbf{y} \in \mathbb{U} + \mathbb{W}$. Then $\mathbf{x}, \mathbf{y}$ can be written as

$$\mathbf{x} = \mathbf{u}_1 + \mathbf{w}_1, \mathbf{y} = \mathbf{u}_2 + \mathbf{w}_2,$$

where $\mathbf{u}_1, \mathbf{u}_2 \in \mathbb{U}$, $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{W}$. Since $\mathbb{U}$ and $\mathbb{W}$ are subspaces of $\mathbb{V}$, they satisfy (1.1.3). Thus,

$$a\mathbf{u}_1 + b\mathbf{u}_2 \in \mathbb{U}, a\mathbf{w}_1 + b\mathbf{w}_2 \in \mathbb{W},$$

so that

$$ax + by = a(\mathbf{u}_1 + \mathbf{w}_1) + b(\mathbf{u}_2 + \mathbf{w}_2)$$
$$= (a\mathbf{u}_1 + b\mathbf{u}_2) + (a\mathbf{w}_1 + b\mathbf{w}_2) \in \mathbb{U} + \mathbb{W}.$$

This shows that $\mathbb{U} + \mathbb{W}$ is closed under the operations of vector addition and scalar multiplication, and hence, it is a subspace of $\mathbb{V}$. ∎

Exercises

Exercise 1 Convert the following decimal representations of real numbers to fractions:

(a) $a = 2.12$,
(b) $b = 0.999\cdots$,
(c) $c = 3.333\cdots$,
(d) $d = 3.141414\cdots$,
(e) $e = 3.14151515\cdots$,
(f) $f = 1.22022022\cdots$.

Exercise 2 In each of the following, determine which of the two numbers is smaller than the other, and which two are equal:

(a) $\frac{5}{6}$ and $\frac{4}{5}$,
(b) 0.83 and $\frac{5}{6}$,
(c) $0.444\cdots$ and $\frac{4}{9}$,
(d) $0.82999\cdots$ and $\frac{5}{6}$.

Exercise 3 Suppose $x = 0.c_1c_2\cdots c_s c_1 c_2 \cdots c_s c_1 c_2 \cdots c_s \cdots = 0.\overline{c_1c_2\cdots c_s}$. Show that
$$x = \frac{c_1 c_2 \cdots c_s}{10^s - 1}.$$

Exercise 4 Show that the union of two countable sets is also countable.

Exercise 5 Simplify the following complex numbers to the form of $\alpha + i\beta$, where $\alpha, \beta \in \mathbb{R}$, and plot them on the complex plane:

(a) $a = 1 + 2i, -a, \bar{a}, -\bar{a}, a + \bar{a}, a - \bar{a}$,
(b) $b = 3 + 4i, \bar{b}, \frac{1}{b}, -b, -\bar{b}$,
(c) $c = 5 + 12i, |c|, |c|^2 - \operatorname{Re} c, |c|^2 + \operatorname{Im} c, \frac{1}{c}$.

Exercise 6 Follow the procedure described below to show that $\sqrt{2} \notin \mathbb{Q}$:

Step 1. Assume $\sqrt{2} \in \mathbb{Q}$. Then $\sqrt{2} = a/b$, where a and b are positive integers with no common divisor different from 1. Verify that a^2 is an even integer.
Step 2. Show that if a^2 is an even integer, then a is also an even integer.
Step 3. Apply the result from *Step 2* to show that b^2 is an even integer.
Step 4. Conclude from *Step 2* that b is also an even integer.

Step 5. Apply the results from *Step 2* and *Step 4* to conclude that a and b have a common divisor (namely 2) different from 1.

Step 6. Conclude that $\sqrt{2}$ is not a rational number, since the assumption in *Step 1* is violated.

Exercise 7 Modify the procedure described in Exercise 6 to show that for any prime number p, its square-root $\sqrt{p}$ is not a rational number.

Exercise 8 Verify that the subset

$$\mathbb{W} = \left\{ (a, b, c) \in \mathbb{R}^3 : a + c = 0 \right\}$$

of $\mathbb{R}^3$ is a subspace of the vector space $\mathbb{R}^3$.

Exercise 9 Justify that the set

$$\mathbb{W} = \left\{ (a, b, c) \in \mathbb{R}^3 : a + c = 1 \right\}$$

is not a subspace of the vector space $\mathbb{R}^3$.

Exercise 10 Verify that the subset

$$\mathbb{W} = \left\{ \begin{bmatrix} a & b & 0 \\ c & 0 & d \end{bmatrix} : a + b = d \right\}$$

of $\mathbb{R}^{2,3}$ is a subspace of the vector space $\mathbb{R}^{2,3}$.

Exercise 11 Justify that the subset

$$\mathbb{W} = \left\{ \begin{bmatrix} a & b & 0 \\ c & 0 & d \end{bmatrix} : a + b = d + 1 \right\}$$

of $\mathbb{R}^{2,3}$ is not a subspace of the vector space $\mathbb{R}^{2,3}$.

Exercise 12 Verify that the subset

$$\mathbb{W} = \left\{ A = [a_{jk}] \in \mathbb{C}^{n,n} : a_{11} + a_{22} + \cdots a_{nn} = 0 \right\}$$

of $\mathbb{C}^{n,n}$ is a subspace of the vector space $\mathbb{C}^{n,n}$.

Exercise 13 Let $1 \le m < n$ be integers, and extend each $\mathbf{x} = (x_1, \ldots, x_m) \in \mathbb{R}^m$ to $\widetilde{\mathbf{x}} = (x_1, \ldots, x_m, 0, \ldots, 0) \in \mathbb{R}^n$ by tacking on $n - m$ zeros to $\mathbf{x}$. If $\widetilde{\mathbf{x}}$ is identified as $\mathbf{x}$, namely $\mathbb{R}^m$ is identified as $\{(x_1, \ldots, x_m, 0, \ldots, 0) : x_j \in \mathbb{R}\}$, show that $\mathbb{R}^m$ is a subspace of $\mathbb{R}^n$. Similarly, fill in the details to identify and conclude that $\mathbb{C}^m$ is a subspace of $\mathbb{C}^n$.

Exercise 14 Let $\mathbb{U}$ and $\mathbb{W}$ be two subspaces of a vector space $\mathbb{V}$. Show that $\mathbb{U} \cap \mathbb{W}$ is a subspace of $\mathbb{V}$.

1.2 Sequence and Function Spaces

In the first half of this section, the Euclidean space $\mathbb{C}^n$ is extended to the vector spaces of infinite (or more precisely, bi-infinite) sequences $\mathbf{x} = \{x_j\}$, where $x_j \in \mathbb{C}$ and j runs from $-\infty$ to ∞. The definitions and results to be presented in this section remain valid for (one-sided) infinite sequences $\mathbf{x} = \{x_j\}$, where j runs from any integer, such as 0, to ∞ (or from $-\infty$ to any integer). We will study the sequence spaces ℓ_p, where $0 \leq p \leq \infty$. For this purpose, we quantify the infinite sequences $\mathbf{x} = \{x_j\}$, by introducing the following measurements $\|\mathbf{x}\|_p$ of $\mathbf{x}$.

(a) For $1 \leq p < \infty$,

$$\|\mathbf{x}\|_p = \left(\sum_{j=-\infty}^{\infty} |x_j|^p \right)^{1/p}. \tag{1.2.1}$$

(b) For $p = \infty$,

$$\|\mathbf{x}\|_\infty = \sup_j |x_j|, \tag{1.2.2}$$

where $\sup_j |x_j|$ (with sup standing for the word "supremum") is the least upper bound of a bounded sequence $\mathbf{x}$, and $\sup_j |x_j| = \infty$ for unbounded $\mathbf{x}$. Namely, $\sup_j |x_j|$ denotes the smallest M such that

$$|x_j| \leq M, \quad \text{for all } j = \ldots, -1, 0, 1, \ldots.$$

(c) For $0 < p < 1$,

$$\|\mathbf{x}\|_p = \sum_{j=-\infty}^{\infty} |x_j|^p. \tag{1.2.3}$$

(d) For $p = 0$,

$$\|\mathbf{x}\|_0 = \# \text{ of non-zero terms in } \mathbf{x}. \tag{1.2.4}$$

By using the above measurements for infinite sequences, we now introduce the following subsets of the set of infinite sequences of complex numbers:

$$\ell_p = \big\{ \mathbf{x} = \{x_j\} : \|\mathbf{x}\|_p < \infty, x_j \in \mathbb{C} \big\}.$$

When there is no ambiguity, the same notation ℓ_p will be used for the subset of infinite sequences $\mathbf{x}$ of real numbers for each p, where $0 \leq p \leq \infty$.

It follows from the above definition of measurements that ℓ_∞ is precisely the set of all bounded infinite sequences, and ℓ_0 is the set of all infinite sequences, each of which has at most finitely many non-zero terms. We remark that the ℓ_0-measurement in (1.2.4) is important to the study of "sparse data representation", and particularly to the concept of "compressed sensing".

Theorem 1 | **Vector space ℓ_p** | *With the operations of vector addition and scalar multiplication defined analogously as those for $\mathbb{C}^n$, namely:*

$$\{x_j\} + \{y_j\} = \{x_j + y_j\}$$

and

$$c\{x_j\} = \{cx_j\}$$

for any $c \in \mathbb{C}$, the set ℓ_p, where $0 \le p \le \infty$, is a vector space over the scalar field $\mathbb{C}$. The same statement is valid for infinite sequences of real numbers over the scalar field $\mathbb{R}$.

Since ℓ_p is a subset of the set of all infinite sequences and since the closure condition on scalar multiplication of sequences in ℓ_p is obvious, it is sufficient to prove the closure property of vector addition, which is a consequence of the following result to be called "triangle inequality" for sequences in ℓ_p.

Theorem 2 | **Triangle inequality for ℓ_p** | *For each real number p, $0 \le p \le \infty$, the ℓ_p-measurement $\|\cdot\|_p$ for the set ℓ_p, as defined by (1.2.1)–(1.2.4), has the following property:*

$$\|\mathbf{x} + \mathbf{y}\|_p \le \|\mathbf{x}\|_p + \|\mathbf{y}\|_p, \tag{1.2.5}$$

for all sequences $\mathbf{x}, \mathbf{y} \in \ell_p$.

The proof of (1.2.5) is simple for $p = 0, 1, \infty$. For example, for $p = 0$, since the number of non-zero terms of the sequence $\mathbf{x} + \mathbf{y}$ does not exceed the sum of the number of non-zero terms of $\mathbf{x}$ and that of $\mathbf{y}$, for $\mathbf{x}, \mathbf{y} \in \ell_0$, the triangle inequality (1.2.5) holds for $p = 0$. The verification for $p = \infty$ is left to the reader as an exercise. For $0 < p < \infty$, the triangle inequality (1.2.5) can be written out precisely as follows: For $0 < p < 1$,

$$\sum_j |x_j + y_j|^p \le \sum_j |x_j|^p + \sum_j |y_j|^p; \tag{1.2.6}$$

and for $1 \le p < \infty$,

$$\left(\sum_j |x_j + y_j|^p\right)^{1/p} \le \left(\sum_j |x_j|^p\right)^{1/p} + \left(\sum_j |y_j|^p\right)^{1/p}. \tag{1.2.7}$$

To prove (1.2.6), it is sufficient to show that if $0 < p < 1$, then

$$|a + b|^p \le |a|^p + |b|^p \tag{1.2.8}$$

for all complex numbers a and b, since (1.2.6) follows by summing both sides of the inequality $|x_j + y_j|^p \le |x_j|^p + |y_j|^p$ over all j.

To prove (1.2.8), we may assume that $a \ne 0$, so that (1.2.8) becomes

$$|1 + x|^p \le 1 + |x|^p, \tag{1.2.9}$$

by setting $b = ax$ and cancelling $|a|^p$. Let us first assume that x is real. In this case, the inequality (1.2.9) trivially holds for $x < 0$, since $|1+x|^p < \max(1, |x|^p) < 1+|x|^p$. For $x \ge 0$, we consider the function

$$\begin{aligned} f(x) &= 1 + |x|^p - |1 + x|^p \\ &= 1 + x^p - (1 + x)^p, \end{aligned}$$

with derivative given by

$$\begin{aligned} f'(x) &= px^{p-1} - p(1 + x)^{p-1} \\ &= p\left(\frac{1}{x^{1-p}} - \frac{1}{(1+x)^{1-p}}\right). \end{aligned}$$

But $x < 1 + x$ is equivalent to

$$\frac{1}{x} > \frac{1}{1 + x},$$

so that since $1 - p > 0$, we have

$$\frac{1}{x^{1-p}} > \frac{1}{(1+x)^{1-p}},$$

yielding $f'(x) > 0$ for all $x > 0$. Hence, since $f(0) = 0$, we may conclude that $f(x) \ge 0$ for all $x \ge 0$, which is equivalent to (1.2.9). In general, suppose that x is complex. Then since $|1 + x| \le 1 + |x|$, the inequality (1.2.9) for real $|x|$ can be applied to prove that (1.2.9) remains valid for complex x (see Exercise 10).   ∎

Inequality (1.2.7) is called Minkowski's inequality for sequences. Since (1.2.7) trivially holds for $p = 1$, we only consider $1 < p < \infty$, and will derive (1.2.7) for $1 < p < \infty$ by applying Hölder's inequality (to be stated later). But let us first establish the following inequality of Young.

Theorem 3 $\boxed{\textbf{Young's inequality}}$ *Let* $1 < p < \infty$ *and* $q = \frac{p}{p-1}$, *called the conjugate of* p. *That is, the pair* $\{p, q\}$ *satisfies the duality condition:*
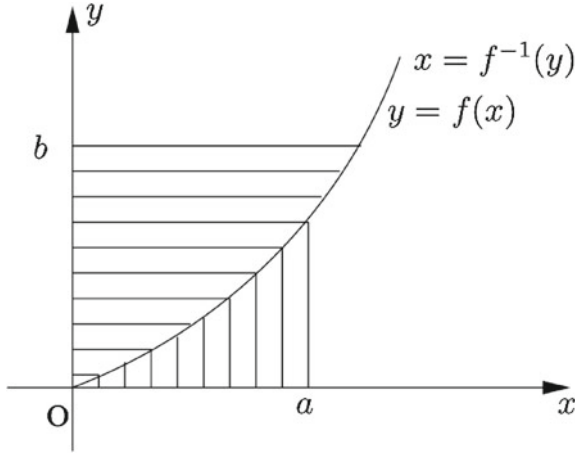
Fig. 1.2   Areas under $y = f(x)$ and $x = f^{-1}(y)$ in the proof of Young's inequality

$$\frac{1}{p} + \frac{1}{q} = 1. \tag{1.2.10}$$

Then for all nonnegative real numbers a and b,

$$ab \le \frac{a^p}{p} + \frac{b^q}{q}, \tag{1.2.11}$$

where equality holds if and only if $b = a^{p-1}$.

Proof   Young's inequality (1.2.11) can be proved as follows. Let $f(x) = x^{p-1}$, for $x \in [0, \infty)$. Then $y = f(x)$ is an increasing function of x for $x \ge 0$, with $f(0) = 0$ and the inverse of f given by $x = f^{-1}(y) = y^{\frac{1}{p-1}}$. Now, it is clear from Fig. 1.2 that

$$\int_0^a f(x)dx + \int_0^b f^{-1}(y)dy \ge ab,$$

where equality holds if and only if $b = f(a) = a^{p-1}$. By direct calculation, we have

$$\int_0^a f(x)dx = \int_0^a x^{p-1}dx = \frac{a^p}{p},$$

and

$$\int_0^b f^{-1}(y)dy = \int_0^b y^{\frac{1}{p-1}}dy = \frac{b^{\frac{1}{p-1}+1}}{\frac{1}{p-1}+1} = \frac{b^q}{q},$$

since

$$\frac{1}{p-1} + 1 = \frac{p}{p-1} = \frac{1}{1-\frac{1}{p}} = q.$$

This completes the proof of Theorem 3.   ■

We are now ready to state and prove the following inequality of Hölder.

Theorem 4 | **Hölder′s inequality for sequences** | *For any real number p with $1 < p < \infty$, let q be its conjugate as defined by* (1.2.10). *Then for any two sequences* $\mathbf{x} = \{x_j\} \in \ell_p$ *and* $\mathbf{y} = \{y_j\} \in \ell_q$, *the sequence* $\{x_j y_j\}$ *is a sequence in* ℓ_1, *and*

$$\sum_j |x_j y_j| \leq \left(\sum_j |x_j|^p \right)^{1/p} \left(\sum_j |y_j|^q \right)^{1/q}. \qquad (1.2.12)$$

Proof To prove this theorem, we first observe that if at least one of the two sequences $\{x_j\}$ and $\{y_j\}$ is the zero sequence (that is, all terms of the sequence are equal to zero), then (1.2.12) trivially holds, since both sides of (1.2.12) would be equal to zero. Hence, we may assume that

$$\|\mathbf{x}\|_p = \left(\sum_j |x_j|^p \right)^{\frac{1}{p}} \neq 0, \ \|\mathbf{y}\|_q = \left(\sum_j |y_j|^q \right)^{\frac{1}{q}} \neq 0,$$

so that (1.2.12) can be written as

$$\sum_j \frac{|x_j|}{\|\mathbf{x}\|_p} \frac{|y_j|}{\|\mathbf{y}\|_q} \leq 1.$$

Now, since

$$\left\| \frac{\mathbf{x}}{\|\mathbf{x}\|_p} \right\|_p = 1, \ \left\| \frac{\mathbf{y}}{\|\mathbf{y}\|_q} \right\|_q = 1,$$

it is sufficient to prove (1.2.12) for $\|\mathbf{x}\|_p = 1$ and $\|\mathbf{y}\|_q = 1$, or equivalently $\sum_j |x_j|^p = 1$ and $\sum_j |y_j|^q = 1$. To do so, we apply Young's inequality to yield

$$|x_j y_j| \leq \frac{|x_j|^p}{p} + \frac{|y_j|^q}{q}.$$

Then, by taking summation of both sides, we have

$$\sum_j |x_j y_j| \leq \frac{\sum_j |x_j|^p}{p} + \frac{\sum_j |y_j|^q}{q} = \frac{1}{p} + \frac{1}{q} = 1.$$

This completes the proof of (1.2.12). ∎

We remark that the special case of $p = q = 2$ in (1.2.12) is called the Cauchy-Schwarz inequality, to be derived in the next section for the general inner product setting.

We are now ready to derive Minkowski's inequality (1.2.7) by applying Hölder's inequality (1.2.12).

To prove (1.2.7), let us first observe that it trivially holds for $p = 1$. So, let us consider p for $1 < p < \infty$ and its conjugate q, as defined by (1.2.10). Then by applying (1.2.12), we have

$$
\begin{aligned}
\sum_j |x_j + y_j|^p &= \sum_j |x_j + y_j||x_j + y_j|^{p-1} \\
&\leq \sum_j |x_j||x_j + y_j|^{p-1} + \sum_j |y_j||x_j + y_j|^{p-1} \\
&\leq \left(\sum_j |x_j|^p\right)^{1/p}\left(\sum_j |x_j + y_j|^{(p-1)q}\right)^{1/q} \\
&\quad + \left(\sum_j |y_j|^p\right)^{1/p}\left(\sum_j |x_j + y_j|^{(p-1)q}\right)^{1/q} \\
&= \left(\|\mathbf{x}\|_p + \|\mathbf{y}\|_p\right)\left(\sum_j |x_j + y_j|^p\right)^{1/q},
\end{aligned}
$$

in view of $(p-1)q = p$. This yields (1.2.7) when both sides are divided by $\left(\sum_j |x_j + y_j|^p\right)^{1/q}$, again by applying of (1.2.10). ∎

Example 1 Let $\{k_i\}$ be a finite sequence of (integer) indices. Then for each $p, 0 \leq p \leq \infty$, the set

$$
S_p = \left\{\mathbf{x} = \{x_j\} \in \ell_p : \sum_i x_{k_i} = 0\right\} \tag{1.2.13}
$$

is a subspace of ℓ_p.

Solution Let $\mathbf{x}, \mathbf{y} \in S_p$ and $\mathbf{z} = a\mathbf{x} + b\mathbf{y}$, where $a, b \in \mathbb{C}$. Then since

$$
\sum_i x_{k_i} = 0 \quad \text{and} \quad \sum_i y_{k_i} = 0,
$$

we have

$$
\sum_i z_{k_i} = \sum_i (ax_{k_i} + by_{k_i}) = a\sum_i x_{k_i} + b\sum_i y_{k_i} = a \times 0 + b \times 0 = 0.
$$

Hence, $\mathbf{z} \in S_p$ and we conclude that S_p is a subspace of ℓ_p. ∎

Let us now turn to the study of vector spaces of functions. For this elementary textbook, we only consider functions that are either continuous or piecewise continuous, although a brief discussion of extension to "measurable" functions will be given in the last section, Sect. 1.5, of this chapter.

For a bounded interval J, we say that a function f is **piecewise continuous** on J if f is continuous on J, with the exception of at most a finite number of points on J at which f has finite jumps, meaning that both left-hand limit $f(x_0-)$ and right-hand limit $f(x_0+)$ exist and are finite for each x_0 in J. If $f(x_0-) \neq f(x_0+)$, then x_0 is called a discontinuity of f, and the value $(f(x_0+) - f(x_0-))$ is called the jump of $f(x)$ at $x = x_0$. If J is an unbounded interval, we say f is piecewise continuous on J, if f is piecewise continuous on any bounded subinterval of J.

Let $-\infty < a < b < \infty$. Then the interval J under consideration may be one of the following intervals:

$$J = (a, b), [a, b], (a, b], [a, b), (-\infty, b), (-\infty, b], (a, \infty), \text{ or } [a, \infty). \quad (1.2.14)$$

In the following, we introduce the necessary notations for the sets of functions defined on J to be studied in this book.

(i) $C(J)$ denotes the set of functions continuous on J.
(ii) $PC(J)$ denotes the set of piecewise continuous functions on J.
(iii) $\widetilde{L}_0(J) = \{f \in PC(J) : f(x) = 0, \text{ if } x \notin [c, d] \subset J, \text{ for some } c < d\}$.

It is easy to see that under the standard operations of addition of two functions and multiplication of a function by a constant, the sets $C(J)$, $PC(J)$ and $\widetilde{L}_0(J)$ are vector spaces over the scalar field $\mathbb{F} = \mathbb{C}, \mathbb{R}$ or $\mathbb{Q}$. Next we study the function classes $\widetilde{L}_p(J) \subset PC(J)$ for $0 < p \leq \infty$, by introducing the following measurements:

(a) For $1 \leq p < \infty$,

$$\|f\|_p = \left(\int_J |f|^p \right)^{1/p}. \quad (1.2.15)$$

(b) For $p = \infty$,

$$\|f\|_\infty = \operatorname*{ess\,sup}_x |f(x)|. \quad (1.2.16)$$

(c) For $0 < p < 1$,

$$\|f\|_p = \int_J |f|^p. \quad (1.2.17)$$

In (1.2.16), "ess sup" (which reads: "essential supremum") means that the values of f at discontinuities are ignored in taking the least upper bound. Also, it should be understood that the notation:

$$f(x) = 0, x \in J$$

is used for functions $f(x) = 0$ for any $x \in J$ at which $f(x)$ is continuous (by ignoring the values at the discontinuities). In this regard, we remark that when the class of piecewise continuous functions is extended to measurable functions (to be discussed in Sect. 1.5), then "ess sup" means that the values of f on a subset of J with measure zero is ignored in taking the least upper bound (see the notions of "almost all" and "almost everywhere" with abbreviation "a.e." in Sect. 1.1 on p.3).

Example 2 Let $J = [0, 1]$. Define f and g in $PC(J)$ as follows.

$$f(x) = \begin{cases} 1, & \text{for } 0 < x < 1, \\ 2, & \text{for } x = 0 \text{ and } x = 1; \end{cases}$$

$$g(x) = \begin{cases} -2x, & \text{for } 0 \le x < 1, \\ 3, & \text{for } x = 1. \end{cases}$$

Determine $\|f\|_\infty$ and $\|g\|_\infty$.

Solution By ignoring the values of $f(x)$ at $x = 0$ and $x = 1$, $|f(x)|$ is bounded by 1, namely $|f(x)| \le 1$ for $0 < x < 1$. In addition, it is clear 1 is the smallest upper bound. Thus $\|f\|_\infty = 1$.

Ignoring the value of $g(x)$ at $x = 1$, we see that $|g(x)| = 2x$ is bounded by 2 on $[0, 1)$. Since $\lim_{x \to 1^-} |g(x)| = 2$, 2 is the smallest upper bound. Thus $\|g\|_\infty = 2$. ∎

We are now ready to introduce the function classes $\widetilde{L}_p(J)$, for all real numbers p with $0 < p \le \infty$, as follows:

$$\widetilde{L}_p(J) = \{f \in PC(J) : \|f\|_p < \infty\}.$$

Example 3 Let f, g and h be functions defined on the interval $J = [0, \infty)$ as follows.

$$f(x) = e^{-x}, x \ge 0;$$

$$g(x) = \begin{cases} \frac{1}{\sqrt{x}}, & \text{for } 0 < x \le 1, \\ 0, & \text{for } x = 0 \text{ and } x > 1; \end{cases}$$

$$h(x) = \begin{cases} x, & \text{for } 0 \le x \le 1, \\ 0, & \text{for } x > 1. \end{cases}$$

Verify that

$f \in C(J)$ and $f \in \widetilde{L}_p(J)$ for any $0 < p \le \infty$, but $f \notin \widetilde{L}_0(J)$;
$g \notin PC(J)$, and hence $g \notin \widetilde{L}_p(J)$ for any $0 \le p \le \infty$;
$h \in \widetilde{L}_p(J)$ for any $0 \le p \le \infty$.

Solution The claim for the function f should be easy to verify by direct computation. To verify the claim for the function g, observe that g has an infinite jump at $x = 0$, since $\lim_{x \to 0+} g(x)$ does not exist. Thus, $g \notin PC(J)$ and therefore not in $\widetilde{L}_p(J)$ for

any $0 \leq p \leq \infty$, although the improper integral $\int_J |g(x)|^p dx < \infty$ for $0 < p < 2$. As to the function h, it is clear that $h \in \widetilde{L}_p(J)$ for any $0 < p \leq \infty$ by direct computation. To see that h is also in $\widetilde{L}_0(J)$, we simply choose $[c, d] = [0, 1]$, so that $h(x) = 0$ for $x \notin [c, d]$. Thus, by the definition of $\widetilde{L}_0(J)$, we may conclude that $h \in \widetilde{L}_0(J)$.   ∎

Theorem 5   ⏐ **Vector spaces $\widetilde{L}_p$** ⏐   *For each p, $0 < p \leq \infty$, the collection of functions $\widetilde{L}_p(J)$ is a vector space over the scalar field $\mathbb{F} = \mathbb{C}$ or $\mathbb{R}$.*

To prove this theorem, it is sufficient to derive the additive closure property of $\widetilde{L}_p(J)$, by showing that the sum of any two functions in $\widetilde{L}_p(J)$ remains to be in $\widetilde{L}_p(J)$. This follows from the triangle inequality to be established in the following theorem.

Theorem 6   ⏐ **Triangle inequality** ⏐ **for L_p**   *Let $\| \cdot \|_p$ be the $\widetilde{L}_p$-measurement for functions defined by* (1.2.15) –(1.2.17). *Then for any functions $f, g \in \widetilde{L}_p(J)$,*

$$\|f + g\|_p \leq \|f\|_p + \|g\|_p. \tag{1.2.18}$$

The proof of (1.2.18) for $p = \infty$ is simple (see Exercise 14). As to $0 < p < \infty$, we give the proof of (1.2.18) by considering $0 < p < 1$ and $1 \leq p < \infty$ separately, as follows.

For $0 < p < 1$,

$$\int_J |f + g|^p \leq \int_J |f|^p + \int_J |g|^p; \tag{1.2.19}$$

and for $1 \leq p < \infty$,

$$\left(\int_J |f + g|^p \right)^{1/p} \leq \left(\int_J |f|^p \right)^{1/p} + \left(\int_J |g|^p \right)^{1/p}. \tag{1.2.20}$$

The proof of (1.2.19) follows immediately from the inequality (1.2.8) with $a = f(x)$ and $b = g(x)$, and simply by integrating both sides of

$$|f(x) + g(x)|^p \leq |f(x)|^p + |g(x)|^p$$

over the interval J.

Inequality (1.2.20) is called Minkowski's inequality. Since the proof of (1.2.20) is trivial for $p = 1$, we will only consider $1 < p < \infty$ in the following discussions. Also, analogous to the proof of Minkowski's inequality for sequences, the inequality (1.2.20) is a consequence of Hölder's inequality, in the following theorem.

Theorem 7   ⏐ **Hölder′s inequality** ⏐   *Let p be any real number with $1 < p < \infty$ and q be its conjugate, as defined in* (1.2.10). *Then for any $f \in \widetilde{L}_p(J)$ and $g \in \widetilde{L}_q(J)$,*

their product fg is in $\widetilde{L}_1(J)$, and

$$\int_J |fg| \leq \left(\int_J |f|^p \right)^{1/p} \left(\int_J |g|^q \right)^{1/q}. \tag{1.2.21}$$

The special case of (1.2.21), where $p = q = 2$, is called the Cauchy-Schwarz inequality, which is an example of the Cauchy-Schwarz inequality for the general inner product setting to be derived in the next section. The proof of (1.2.21) is provided in the following.

First, dividing f and g by $\left(\int_J |f|^p \right)^{\frac{1}{p}}$ and $\left(\int_J |f|^q \right)^{\frac{1}{q}}$ respectively, we may assume that

$$\int_J |f|^p = \int_J |g|^q = 1.$$

Next, by Young's inequality (3), we have

$$|fg| \leq \frac{|f|^p}{p} + \frac{|g|^q}{q}.$$

Hence, integration of both sides over the interval J yields

$$\int_J |fg| \leq \frac{\int_J |f|^p}{p} + \frac{\int_J |g|^q}{q} = \frac{1}{p} + \frac{1}{q} = 1,$$

as desired.   ∎

Since the proof of (1.2.20) by applying (1.2.21) is essentially the same as that of (1.2.7) by applying (1.2.12), it is safe to leave the derivation as an exercise (see Exercise 12).

<div align="center">

Exercises

</div>

Exercise 1 Let $\mathbb{W}_a = \{f \in C(-\infty, \infty) : f(0) = a\}$, where $a \in \mathbb{R}$ is fixed. Determine the value of a for which $\mathbb{W}_a$ is a subspace of $C(-\infty, \infty)$. Justify your answer with rigorous reasoning.

Exercise 2 Let $\mathbb{V}_a = \{f \in C[-2, 2] : f(-1) + f(1) = a\}$, where $a \in \mathbb{R}$ is fixed. Determine the value of a for which $\mathbb{V}_a$ is a subspace of $C[-2, 2]$. Justify your answer with rigorous reasoning.

Exercise 3 In each of the following, determine the range of the real number α for which the following infinite sequence is in ℓ_1.

(a) $\mathbf{x} = \{(1 + k^2)^\alpha\}$, $k = \ldots, -1, 0, 1, \cdots$.
(b) $\mathbf{y} = \{2^{\alpha k}\}$, $k = \ldots, -1, 0, 1, \cdots$.
(c) $\mathbf{z} = \{2^{\alpha |k|}\}$, $k = \ldots, -1, 0, 1, \cdots$.

Exercise 4   Repeat Exercise 3 for ℓ_p, where $1 < p \leq \infty$. (Note that α may depend on p.)

Exercise 5   In each of the following, determine the range of the real number α for which the function is in $\widetilde{L}_1(J)$, where the interval J may be different in different problems.

(a) $f(x) = (1 + x^2)^\alpha$, $J = (0, 1)$.
(b) $f(x) = (1 + x^2)^\alpha$, $J = (0, \infty)$.
(c) $g(x) = e^{\alpha x}$, $J = (0, \infty)$.

Exercise 6   Repeat Exercise 5 for $\widetilde{L}_p(J)$, where $1 < p \leq \infty$. (Note that α may depend on p.)

Exercise 7   Let $\mathbb{W} = \{\mathbf{x} = \{x_j\} \in \ell_p : \sum_j x_{2j} = 0\}$. Show that $\mathbb{W}$ is a subspace of

$\ell_p, 0 < p < \infty$.

Exercise 8   Let $\mathbb{U} = \{\mathbf{x} = \{x_j\} \in \ell_p : \sum_j (-1)^j x_j = 0\}$. Is $\mathbb{U}$ a subspace of ℓ_p?
Justify your answer with rigorous reasoning.

Exercise 9   Let $J_1 \subset J$ and $\mathbb{W}$ be a subset of $\widetilde{L}_p(J), 0 < p < \infty$, where

$$\int_{J_1} f = 0, \text{ for all } f \in \mathbb{W}.$$

Is $\mathbb{W}$ a subspace of $\widetilde{L}_p(J)$? Justify your answer with rigorous reasoning.

Exercise 10   Extend the inequality (1.2.9) to the complex setting. That is, for $0 < p < 1$, show that $|1 + z|^p \leq 1 + |z|^p$ for all $z \in \mathbb{C}$.
   Hint: Use the facts that $|1 + z| \leq 1 + |z|$ and that $f(x) = x^p$ is increasing for $x > 0$. Then apply (1.2.9) to $(1 + |z|)^p$.

Exercise 11   Let $p > 1$ and consider the function $f(x) = 1 + x^p - (1 + x)^p$. Follow the derivation of (1.2.9) to show that $f(x) < 0$ for all $x > 0$. Conclude that the inequality (1.2.9) is reversed if $0 < p \leq 1$ is replaced by $p > 1$.

Exercise 12   Follow the derivation of (1.2.7) and apply (1.2.21) to derive the inequality (1.2.20).

Exercise 13   Prove (1.2.5) for $p = \infty$.

Exercise 14   Prove that $C(J), PC(J), \widetilde{L}_0(J)$ and $\widetilde{L}_\infty(J)$ are vector spaces on $\mathbb{C}$.

1.3 Inner-Product Spaces

In this section, the notions of the "product" of two vectors and the "angle" between them will be introduced. More precisely, the "product" of two vectors (called "inner product") is some scalar and the angle between them is defined in terms of their inner product, after the vectors are normalized.

Definition 1 | **Inner-product space** | *Let $\mathbb{V}$ be a vector space over some scalar field $\mathbb{F}$, which is either $\mathbb{C}$ or any subfield of $\mathbb{C}$, such as the field $\mathbb{R}$ of real numbers. Then $\mathbb{V}$ is called an inner-product space, if there is a function $\langle \cdot, \cdot \rangle$ defined on $\mathbb{V} \times \mathbb{V}$ with range in $\mathbb{F}$, with the following properties :*

(a) Conjugate symmetry: $\langle \mathbf{x}, \mathbf{y} \rangle = \overline{\langle \mathbf{y}, \mathbf{x} \rangle}$ *for all* $\mathbf{x}, \mathbf{y} \in \mathbb{V}$;
(b) Linearity: $\langle a\mathbf{x} + b\mathbf{y}, \mathbf{z} \rangle = a\langle \mathbf{x}, \mathbf{z} \rangle + b\langle \mathbf{y}, \mathbf{z} \rangle$ *for all* $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{V}$, $a, b \in \mathbb{F}$;
(c) Positivity: $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0$ *for all* $\mathbf{x} \in \mathbb{V}$, *and* $\langle \mathbf{x}, \mathbf{x} \rangle = 0 \Leftrightarrow \mathbf{x} = \mathbf{0}$.

From the definition of $\langle \cdot, \cdot \rangle$, we have

$$\langle a\mathbf{x}, \mathbf{y} \rangle = a\langle \mathbf{x}, \mathbf{y} \rangle, \ \langle \mathbf{x}, a\mathbf{y} \rangle = \overline{a}\langle \mathbf{x}, \mathbf{y} \rangle.$$

The function $\langle \cdot, \cdot \rangle$ from $\mathbb{V} \times \mathbb{V}$ to $\mathbb{F}$ is called an **inner product**, also called **scalar product** since its range is the scalar field $\mathbb{F}$. Recall from an elementary course in Linear Algebra that for $\mathbb{V} = \mathbb{R}^n$ and $\mathbb{F} = \mathbb{R}$, the function $\langle \cdot, \cdot \rangle$ defined on $\mathbb{R}^n \times \mathbb{R}^n$ by

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x} \cdot \mathbf{y} = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n, \tag{1.3.1}$$

for all $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ in $\mathbb{R}^n$ is called the "dot product" between $\mathbf{x}$ and $\mathbf{y}$. For the complex field $\mathbb{C}$, we may extend the "dot product" to vectors in $\mathbb{V} = \mathbb{C}^n$, by defining

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x} \cdot \overline{\mathbf{y}} = x_1 \overline{y_1} + x_2 \overline{y_2} + \cdots + x_n \overline{y_n}, \tag{1.3.2}$$

for all $\mathbf{x} = (x_1, \ldots, x_n), \mathbf{y} = (y_1, \ldots, y_n) \in \mathbb{C}^n$, to satisfy the conjugate symmetry condition in the above definition of the inner product.

Although there are other ways to define various inner products for the vector spaces $\mathbb{R}^n$ and $\mathbb{C}^n$, the dot products in (1.3.1) and (1.3.2) are called the standard (or ordinary) inner products for $\mathbb{R}^n$ and $\mathbb{C}^n$. In this book, unless specified otherwise, the inner product for $\mathbb{R}^n$ and $\mathbb{C}^n$ is always the standard inner product, as defined by the dot product in (1.3.1) and (1.3.2).

Example 1 Let $\mathbb{V} = \mathbb{C}^n$ and $\mathbb{F} = \mathbb{C}$. Then the function $\langle \cdot, \cdot \rangle$ defined by (1.3.2) is an inner product for $\mathbb{C}^n$.

Solution We verify (a)–(c) of Definition 1, as follows:

(a) For $\mathbf{x}, \mathbf{y} \in \mathbb{C}^n$,

$$\langle \mathbf{x}, \mathbf{y} \rangle = x_1\bar{y}_1 + \cdots + x_n\bar{y}_n$$
$$= \bar{y}_1x_1 + \cdots + \bar{y}_nx_n$$
$$= \overline{y_1\bar{x}_1 + \cdots + y_n\bar{x}_n} = \overline{\langle \mathbf{y}, \mathbf{x} \rangle}.$$

(b) For all $a, b \in \mathbb{C}$ and $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{C}^n$,

$$\langle a\mathbf{x} + b\mathbf{y}, \mathbf{z} \rangle = (ax_1 + by_1)\bar{z}_1 + \cdots + (ax_n + by_n)\bar{z}_n$$
$$= (ax_1\bar{z}_1 + \cdots + ax_n\bar{z}_n) + (by_1\bar{z}_1 + \cdots + by_n\bar{z}_n)$$
$$= a\langle \mathbf{x}, \mathbf{z} \rangle + b\langle \mathbf{y}, \mathbf{z} \rangle.$$

(c) $\langle \mathbf{x}, \mathbf{x} \rangle = |x_1|^2 + \cdots + |x_n|^2 \geq 0$ and $\langle \mathbf{x}, \mathbf{x} \rangle = 0$ if and only if $|x_1|^2 + \cdots + |x_n|^2 = 0$, or $x_1 = 0, \cdots, x_n = 0$, or $\mathbf{x} = \mathbf{0}$.   ■

Since the proofs of the following two theorems are straight-forward, it is safe to leave them as exercises (see Exercises 2–3).

Theorem 1 | **Inner-product space ℓ_2** | *For the vector space ℓ_2 of infinite sequences over the scalar field $\mathbb{F} = \mathbb{C}$ or $\mathbb{R}$, the function*

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{k=-\infty}^{\infty} x_k\bar{y}_k, \qquad (1.3.3)$$

defined on $\ell_2 \times \ell_2$, is an inner product. Consequently, endowed with this inner product, ℓ_2 is an inner-product space.

Theorem 2 | **Inner-product space $\widetilde{L}_2$** | *For the vector space $\widetilde{L}_2(J)$ over the scalar field $\mathbb{F} = \mathbb{C}$ or $\mathbb{R}$, the function*

$$\langle f, g \rangle = \int_J f(t)\overline{g(t)}dt, \qquad (1.3.4)$$

defined on $\widetilde{L}_2(J) \times \widetilde{L}_2(J)$, is an inner product. Consequently, endowed with this inner product, $\widetilde{L}_2(J)$ is an inner-product space.

It follows from Hölder's inequality for sequences (1.2.12) on p.17 for $p = q = 2$, that

$$\left| \sum_k x_k\bar{y}_k \right| \leq \sum_k |x_k||y_k| \leq \left(\sum_k |x_k|^2 \right)^{\frac{1}{2}} \left(\sum_k |y_k|^2 \right)^{\frac{1}{2}}, \qquad (1.3.5)$$

so that $\langle \mathbf{x}, \mathbf{y} \rangle$ in (1.3.3) is well defined. Similarly, by Hölder's inequality (1.2.21) on p.22 for $p = q = 2$, we also have

$$\left| \int_J f(t)\overline{g(t)}dt \right| \le \int_J |f(t)| \, |g(t)| dt \le \left(\int_J |f(t)|^2 dt \right)^{\frac{1}{2}} \left(\int_J |g(t)|^2 dt \right)^{\frac{1}{2}},$$

(1.3.6)

so that $\langle f, g \rangle$ in (1.3.4) is well defined. The inequalities (1.3.5) and (1.3.6) are also called the Cauchy-Schwarz inequality, to be discussed later in this section.

The inner products defined by (1.3.3) and (1.3.4) are called the standard (or ordinary) inner products for ℓ_2 and $\tilde{L}_2$. In the discussions throughout this book, the inner products for ℓ_2 and $\tilde{L}_2$ are always the standard inner products, unless specified otherwise.

We next define some measurement, called the "norm" (or "length") of a vector in an inner-product space $\mathbb{V}$.

Definition 2 | **Norm for inner-product space** | *Let $\mathbb{V}$ be an inner-product space, with inner product $\langle \cdot, \cdot \rangle$ as in Definition 1. Then*

$$\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle},$$

(1.3.7)

is called the norm (or length) of $\mathbf{x} \in \mathbb{V}$.

The inner product and its corresponding norm as defined above have the following relation.

Theorem 3 | **Cauchy-Schwarz inequality** | *Let $\mathbb{V}$ be an inner-product space over some scalar field $\mathbb{F}$, where $\mathbb{F} = \mathbb{C}$ or a subfield of $\mathbb{C}$. Then for all $\mathbf{x}, \mathbf{y} \in \mathbb{V}$,*

$$|\langle \mathbf{x}, \mathbf{y} \rangle| \le \|\mathbf{x}\| \|\mathbf{y}\|,$$

(1.3.8)

with $\| \cdot \|$ defined by (1.3.7). Furthermore, equality in (1.3.8) holds if and only if

$$\mathbf{x} = c\mathbf{y}, \text{ or } \mathbf{y} = c\mathbf{x}$$

for some scalar (also called constant) $c \in \mathbb{F}$.

Proof We only prove this theorem for $\mathbb{F} = \mathbb{R}$ and leave the proof for $\mathbb{F} = \mathbb{C}$ as an exercise (see Exercise 6). Let $a \in \mathbb{R}$ be any constant. We compute, according to (1.3.7),

$$
\begin{aligned}
0 \le \|\mathbf{x} - a\mathbf{y}\|^2 &= \langle \mathbf{x} - a\mathbf{y}, \mathbf{x} - a\mathbf{y} \rangle \\
&= \langle \mathbf{x} - a\mathbf{y}, \mathbf{x} \rangle + \langle \mathbf{x} - a\mathbf{y}, -a\mathbf{y} \rangle \\
&= \langle \mathbf{x}, \mathbf{x} \rangle - a\langle \mathbf{y}, \mathbf{x} \rangle - a\langle \mathbf{x}, \mathbf{y} \rangle + a^2 \langle \mathbf{y}, \mathbf{y} \rangle \\
&= \langle \mathbf{x}, \mathbf{x} \rangle - 2a\langle \mathbf{x}, \mathbf{y} \rangle + a^2 \langle \mathbf{y}, \mathbf{y} \rangle \\
&= \|\mathbf{x}\|^2 - 2a\langle \mathbf{x}, \mathbf{y} \rangle + a^2 \|\mathbf{y}\|^2.
\end{aligned}
$$

(1.3.9)

If $\mathbf{y} = \mathbf{0}$, then the theorem trivially holds. So we may assume $\mathbf{y} \ne \mathbf{0}$. Then by setting

$$a = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{y}\|^2},$$

we have

$$\begin{aligned} 0 \;&\leq\; \|\mathbf{x}\|^2 - 2\frac{\langle \mathbf{x}, \mathbf{y} \rangle \langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{y}\|^2} + \frac{|\langle \mathbf{x}, \mathbf{y} \rangle|^2 \|\mathbf{y}\|^2}{\|\mathbf{y}\|^4} \\ &= \|\mathbf{x}\|^2 - \frac{|\langle \mathbf{x}, \mathbf{y} \rangle|^2}{\|\mathbf{y}\|^2} \end{aligned}$$

or

$$0 \leq \|\mathbf{x}\|^2 \|\mathbf{y}\|^2 - |\langle \mathbf{x}, \mathbf{y} \rangle|^2,$$

which is the same as (1.3.8). Moreover, the equality in (1.3.8) holds if and only if

$$0 = \|\mathbf{x} - a\mathbf{y}\|^2$$

in (1.3.9) with $a = \dfrac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{y}\|^2}$, or $\mathbf{x} = a\mathbf{y}$.   ∎

Next, observe that for a real inner-product space $\mathbb{V}$ over the field $\mathbb{R}$ of real numbers, with norm defined by (1.3.7), it follows from the Cauchy-Schwarz inequality (1.3.8) that

$$-1 \leq \left\langle \frac{\mathbf{x}}{\|\mathbf{x}\|}, \frac{\mathbf{y}}{\|\mathbf{y}\|} \right\rangle \leq 1$$

for all non-zero $\mathbf{x}, \mathbf{y} \in \mathbb{V}$, since $\langle \mathbf{x}, \mathbf{y} \rangle \in \mathbb{R}$. Hence, there is precisely one value θ, with $0 \leq \theta \leq \pi$ such that

$$\cos\theta = \left\langle \frac{\mathbf{x}}{\|\mathbf{x}\|}, \frac{\mathbf{y}}{\|\mathbf{y}\|} \right\rangle = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\| \|\mathbf{y}\|},$$

or

$$\theta = \cos^{-1}\left(\frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\| \|\mathbf{y}\|} \right), \tag{1.3.10}$$

which is called the "**angle**" between the two non-zero real-valued vectors $\mathbf{x}$ and $\mathbf{y}$.

Remark 1 (a) For $\theta = 0$, $\mathbf{x} = c\mathbf{y}$ for some constant $c > 0$. In other words, $\mathbf{x}$ and $\mathbf{y}$ are parallel and "pointing to the same direction".
(b) For $\theta = \pi$, $\mathbf{x} = c\mathbf{y}$ for some $c < 0$. In other words, $\mathbf{x}$ and $\mathbf{y}$ are parallel, but "pointing to opposite directions".
(c) For $\theta = \pi/2$, $\langle \mathbf{x}, \mathbf{y} \rangle = 0$. In other words, $\mathbf{x}$ and $\mathbf{y}$ are perpendicular to each other.
(d) In view of the previous remark, there is no need to consider the notion of the "angle" between two complex-valued vectors $\mathbf{x}$ and $\mathbf{y}$ to convey that they are perpendicular to each other. All that is needed is that their inner product is zero: $\langle \mathbf{x}, \mathbf{y} \rangle = 0$. In the literature, the term "orthogonality" is used most frequently.∎

Definition 3 | **Orthogonal vectors** | *Let $\mathbb{V}$ be an inner-product space over $\mathbb{C}$ or a sub-field of $\mathbb{C}$. Then $\mathbf{x}, \mathbf{y} \in \mathbb{V}$ are said to be orthogonal to each other, if $\langle \mathbf{x}, \mathbf{y} \rangle = 0$.*

Observe that since $\langle \mathbf{x}, \mathbf{0} \rangle = 0$ for all $\mathbf{x} \in \mathbb{V}$, the zero vector $\mathbf{0}$ is orthogonal to all vectors in the inner-product space $\mathbb{V}$.

For a subspace $\mathbb{W} \subset \mathbb{V}$, the notion $\mathbb{W}^{\perp}$ (which reads $\mathbb{W}$ "perp" for the word "perpendicular") is defined by

$$\mathbb{W}^{\perp} = \big\{ \mathbf{x} \in \mathbb{V} : \langle \mathbf{x}, \mathbf{w} \rangle = 0, \text{ for all } \mathbf{w} \in \mathbb{W} \big\}. \tag{1.3.11}$$

It can be verified that $\mathbb{W}^{\perp}$ is a vector space and $\mathbb{W} \cap \mathbb{W}^{\perp} = \{\mathbf{0}\}$ (see Exercises 13–14). In this book, $\mathbb{W}^{\perp}$ is called the orthogonal complement of $\mathbb{W}$ in $\mathbb{V}$, or the orthogonal complementary subspace of $\mathbb{V}$ relative to $\mathbb{W}$.

Example 2 Let $\mathbf{x} = (1, 1, 4)$, $\mathbf{y} = (2, -1, 2) \in \mathbb{R}^3$. Find the angle between $\mathbf{x}$ and $\mathbf{y}$. Repeat this for $\mathbf{u} = (1, -2, 2)$, $\mathbf{v} = (-2, 1, 2)$.

Solution We have

$$\langle \mathbf{x}, \mathbf{y} \rangle = 1(2) + 1(-1) + 4(2) = 9,$$
$$\|\mathbf{x}\| = \sqrt{1^2 + 1^2 + 4^2} = \sqrt{18} = 3\sqrt{2},$$
$$\|\mathbf{y}\| = \sqrt{2^2 + (-1)^2 + 2^2} = \sqrt{9} = 3.$$

Thus,

$$\cos \theta = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\| \|\mathbf{y}\|} = \frac{9}{3\sqrt{2}(3)} = \frac{\sqrt{2}}{2}.$$

The angle θ between $\mathbf{x}$ and $\mathbf{y}$ is

$$\theta = \cos^{-1} \frac{\sqrt{2}}{2} = \frac{\pi}{4}.$$

For $\mathbf{u}$ and $\mathbf{v}$, we have

$$\langle \mathbf{u}, \mathbf{v} \rangle = 1(-2) + (-2)(1) + 2(2) = 0.$$

Thus, $\mathbf{u}$ and $\mathbf{v}$ are orthogonal to each other, and the angle between them is $\frac{\pi}{2}$. ∎

Example 3 Let $f_1(x) = 1$, $f_2(x) = x$, $x \in [0, 1]$ be the functions in $\widetilde{L}_2[0, 1]$. Find the angle between f_1 and f_2. Repeat this for $g_1(x) = 1$, $g_2(x) = \cos \pi x$.

Solution We have

$$\langle f_1, f_2 \rangle = \int_0^1 f_1(x)\overline{f_2(x)}dx = \int_0^1 xdx = \frac{1}{2},$$

$$\|f_1\|^2 = \int_0^1 |f_1(x)|^2dx = \int_0^1 1dx = 1,$$

$$\|f_2\|^2 = \int_0^2 |f_2(x)|^2dx = \int_0^1 x^2dx = \frac{1}{3};$$

and

$$\cos\theta = \frac{\langle f_1, f_2 \rangle}{\|f_1\|\|f_2\|} = \frac{\sqrt{3}}{2}.$$

Thus, the angle θ between f_1 and f_2 is

$$\theta = \cos^{-1}\frac{\sqrt{3}}{2} = \frac{\pi}{6}.$$

For g_1 and g_2, we have

$$\langle g_1, g_2 \rangle = \int_0^1 \cos\pi x\, dx = \left[\frac{1}{\pi}\sin\pi x\right]_{x=0}^1 = 0.$$

Therefore g_1 and g_2 are orthogonal to each other, and the angle between them is $\frac{\pi}{2}$.

Observe that there is no geometric meaning for the angle between two functions. ∎

Theorem 4 ⎢**Pythagorean theorem**⎢ *Let $\mathbb{V}$ be an inner-product space over some scalar field $\mathbb{F} = \mathbb{C}$ or $\mathbb{R}$. If $\mathbf{x}, \mathbf{y} \in \mathbb{V}$ are orthogonal, then*

$$\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 = \|\mathbf{x} + \mathbf{y}\|^2. \tag{1.3.12}$$

When $\mathbf{x}, \mathbf{y} \in \mathbb{R}^2$, equality (1.3.12) is the Pythagorean theorem from middle/high-school geometry: the square of the length of the hypotenuse of a right-angled triangle is equal to the sum of the squares of the lengths of the other two sides ("legs"). The equality (1.3.12) can be derived as follows:

$$\begin{aligned}
\|\mathbf{x} + \mathbf{y}\|^2 &= \langle \mathbf{x} + \mathbf{y}, \mathbf{x} + \mathbf{y} \rangle \\
&= \langle \mathbf{x}, \mathbf{x} \rangle + \langle \mathbf{x}, \mathbf{y} \rangle + \langle \mathbf{y}, \mathbf{x} \rangle + \langle \mathbf{y}, \mathbf{y} \rangle \\
&= \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2,
\end{aligned}$$

where the last equality follows from the orthogonality assumption: $\langle \mathbf{x}, \mathbf{y} \rangle = 0$, $\langle \mathbf{y}, \mathbf{x} \rangle = 0$. ∎

Example 4 Let $\mathbf{u} = (1, -2, 2)$, $\mathbf{v} = (-2, 1, 2) \in \mathbb{R}^3$ be the vectors considered in Example 2. Verify the Pythagorean theorem by computing $\|\mathbf{u}\|^2$, $\|\mathbf{v}\|^2$ and $\|\mathbf{u} + \mathbf{v}\|^2$ directly.

Solution We have

$$\|\mathbf{u}\|^2 = 1^1 + (-2)^2 + 2^2 = 9,$$
$$\|\mathbf{v}\|^2 = (-2)^2 + 1^2 + 2^2 = 9,$$

and

$$\|\mathbf{u} + \mathbf{v}\|^2 = \|(-1, -1, 4)\|^2 = (-1)^2 + (-1)^2 + 4^2 = 18,$$

which is indeed the sum of $\|\mathbf{u}\|^2$ and $\|\mathbf{v}\|^2$.   ∎

Example 5 Let $g_1(x) = 1$, $g_2(x) = \cos \pi x$ be the functions in $\tilde{L}_2[0, 1]$ considered in Example 3. Verify the Pythagorean theorem with $g_1(x)$ and $g_2(x)$ by computing $\|g_1\|^2$, $\|g_2\|^2$ and $\|g_1 + g_2\|^2$ directly.

Solution We have

$$\|g_1\|^2 = \int_0^1 |g_1(x)|^2 dx = \int_0^1 1^2 dx = 1,$$

$$\|g_2\|^2 = \int_0^1 |g_2(x)|^2 dx = \int_0^1 \cos^2 \pi x dx$$

$$= \int_0^1 \frac{1}{2}(1 + \cos 2\pi x) dx = \frac{1}{2}.$$

On the other hand, we have

$$\|g_1 + g_2\|^2 = \int_0^1 |1 + \cos \pi x|^2 dx$$

$$= \int_0^1 1 + 2 \cos \pi x + \cos^2 \pi x dx$$

$$= 1 + 0 + \frac{1}{2} = \frac{3}{2}.$$

Thus, (1.3.12) holds for $g_1(x)$ and $g_2(x)$.   ∎

A vector $\mathbf{x}$ in a vector space over $\mathbb{F}$ is said to be a finite linear combination of vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, if there exist scalars (also called constants) $c_1, \ldots, c_n \in \mathbb{F}$, such that

$$\mathbf{x} = \sum_{j=1}^n c_j \mathbf{x}_j = c_1 \mathbf{x}_1 + \cdots + c_n \mathbf{x}_n. \tag{1.3.13}$$

This concept is used to introduce the notion of the "span" or more precisely, "linear span" of a set of vectors, as follows.

Definition 4 | **Linear span** | *Let $S = \{\mathbf{x}_j\}$ be a (finite or infinite) set of vectors in a vector space $\mathbb{V}$ over a scalar field $\mathbb{F}$. Then the (linear) span of S, denoted by $\mathrm{span} S$ or $\mathrm{span}\{\mathbf{x}_j\}$, is the set of all finite linear combinations of vectors in S.*

In particular, the span of a finite set $S = \{\mathbf{x}_j : 1 \leq j \leq n\}$ is

$$\mathrm{span} S = \left\{ \mathbf{x} = \sum_{j=1}^{n} c_j \mathbf{x}_j : c_1, \ldots, c_n \in \mathbb{F} \right\}.$$

It is easy to show that $\mathrm{span}\{\mathbf{x}_j\}$ is a subspace of $\mathbb{V}$ (see Exercise 15). Thus, alternatively, the set $\{\mathbf{x}_j\}$ is said to span the vector space: $\mathrm{span}\{\mathbf{x}_j\}$.

Example 6   Let $\mathbf{x}_1 = (1, 1, 3, -5)$, $\mathbf{x}_2 = (2, -1, 2, 4) \in \mathbb{R}^4$ and $\mathbb{W} = \mathrm{span}\{\mathbf{x}_1, \mathbf{x}_2\}$. Find $\mathbb{W}^{\perp}$.

Solution A vector $\mathbf{c} = (c_1, c_2, c_3, c_4) \in \mathbb{R}^4$ is in $\mathbb{W}^{\perp}$ if and only if $\langle \mathbf{c}, \mathbf{x}_1 \rangle = 0$ and $\langle \mathbf{c}, \mathbf{x}_2 \rangle = 0$. That is,

$$\begin{cases} c_1 + c_2 + 3c_3 - 5c_4 = 0, \\ 2c_1 - c_2 + 2c_3 + 4c_4 = 0. \end{cases}$$

Solving the above linear system, we have

$$c_1 = -\frac{5}{3}c_3 + \frac{1}{3}c_4, \quad c_2 = -\frac{4}{3}c_3 + \frac{14}{3}c_4.$$

Thus,

$$\begin{aligned} \mathbf{c} &= \left(-\frac{5}{3}c_3 + \frac{1}{3}c_4, -\frac{4}{3}c_3 + \frac{14}{3}c_4, c_3, c_4 \right) \\ &= c_3 \left(-\frac{5}{3}, -\frac{4}{3}, 1, 0 \right) + c_4 \left(\frac{1}{3}, \frac{14}{3}, 0, 1 \right) \\ &= \frac{c_3}{3}(-5, -4, 3, 0) + \frac{c_4}{3}(1, 14, 0, 3). \end{aligned}$$

Therefore $\mathbb{W}^{\perp} = \mathrm{span}\{(-5, -4, 3, 0), (1, 14, 0, 3)\}$.   ∎

<div align="center">| **Exercises** |</div>

Exercise 1   Let $\langle \cdot, \cdot \rangle$ be the inner product as introduced in Definition 1. Show that $\langle \mathbf{x}, a\mathbf{y} \rangle = \bar{a}\langle \mathbf{x}, \mathbf{y} \rangle$, where $\bar{a}$ is the complex conjugate of $a \in \mathbb{C}$.

Exercise 2   Provide a proof of Theorem 1.

Exercise 3   Provide a proof of Theorem 2.

Exercise 4 In each of the following, compute $\langle \mathbf{x}, \mathbf{y} \rangle$, $\|\mathbf{x}\|$, and $\|\mathbf{y}\|$. Then illustrate the Cauchy-Schwarz inequality (1.3.8) with these examples.

(a) $\mathbf{x}, \mathbf{y} \in \mathbb{R}^3$, with $\mathbf{x} = (1, 2, -1)$ and $\mathbf{y} = (0, -1, 4)$.

(b) $\mathbf{x}, \mathbf{y} \in \ell_2$, with $\mathbf{x} = \{x_j\}$ and $\mathbf{y} = \{y_j\}$, where $x_j = y_j = 0$ for $j \leq 0$;
$x_j = \dfrac{1}{j}$, $y_j = (-1)^j \dfrac{1}{j}$, for $j > 0$.

Hint: $\displaystyle\sum_{j=1}^{\infty} \frac{1}{j^2} = \frac{\pi^2}{6}$. Compute $\displaystyle\sum_{j=1}^{\infty} \frac{1}{(2j-1)^2}$ by observing that

$$\sum_{j=1}^{\infty} \frac{1}{j^2} = \sum_{j=1}^{\infty} \frac{1}{(2j-1)^2} + \sum_{j=1}^{\infty} \frac{1}{(2j)^2}.$$

Apply the same trick of writing the summation as the sum over the even indices and the sum over the odd indices to compute $\displaystyle\sum_{j=1}^{\infty} \frac{(-1)^j}{j^2}$.

Exercise 5 In each of the following, compute $\langle f, g \rangle$, $\|f\|$, and $\|g\|$ for the space $\widetilde{L}_2[0, 1]$. Then illustrate the Cauchy-Schwarz inequality (1.3.8) with these examples.

(a) $f(x) = 1 + x$, $g(x) = 1 - x$.
(b) $f(x) = \sin \pi x$, $g(x) = \cos \pi x$.
(c) $f(x) = 1$, $g(x) = \cos 2\pi x$.
(d) $f(x) = \cos 2\pi m x$, $g(x) = \cos 2\pi n x$, where m and n are integers.

Exercise 6 Extend the proof of Theorem 3 from $\mathbb{F} = \mathbb{R}$ to $\mathbb{F} = \mathbb{C}$.
  Hint: If the scalar $a \in \mathbb{R}$ defined in the proof of the theorem is replaced by $a \in \mathbb{C}$, then $\bar{a} = \langle \mathbf{y}, \mathbf{x} \rangle / \|\mathbf{y}\|^2$.

Exercise 7 As a continuation of Exercise 4, compute the angle θ between the vectors $\mathbf{x}$ and $\mathbf{y}$. (Note that θ is restricted to $0 \leq \theta \leq \pi$.)

Exercise 8 As a continuation of Exercise 5, compute the angle θ between the functions f and g in $\widetilde{L}_2[0, 1]$.

Exercise 9 Let $\mathbb{V}$ be an inner-product space over some scalar field $\mathbb{F}$ and let the norm $\| \cdot \|$ of vectors in $\mathbb{V}$ be defined as in (1.3.7). Derive the following "parallelogram law" for all vectors $\mathbf{x}, \mathbf{y} \in \mathbb{V}$:

$$\|\mathbf{x} + \mathbf{y}\|^2 + \|\mathbf{x} - \mathbf{y}\|^2 = 2\|\mathbf{x}\|^2 + 2\|\mathbf{y}\|^2.$$

Exercise 10 In Exercise 9, consider the special case of $\mathbb{V} = \mathbb{R}^2$ and $\mathbb{F} = \mathbb{R}$. Recall that $\mathbf{x} = [x_1, x_2]^T \in \mathbb{V}$ is also considered as a point $\mathbf{x} = (x_1, x_2)$ on the plane $\mathbb{R}^2$. Let $\mathbf{y} = (y_1, y_2)$ be another point on the plane $\mathbb{R}^2$. Provided that $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{x} + \mathbf{y}$ are different from $\mathbf{0}$, then the four points $\mathbf{0} = (0, 0)$, $\mathbf{x} = (x_1, x_2)$, $\mathbf{y} = (y_1, y_2)$, and

$\mathbf{x} + \mathbf{y} = (x_1 + y_1, x_2 + y_2)$ constitute the four vertices of a parallelogram. Show that the lengths of the two diagonals are given by $\|\mathbf{x} + \mathbf{y}\|$ and $\|\mathbf{x} - \mathbf{y}\|$. Hence, the parallelogram law in Exercise 9 says that the sum of the squares on the two diagonals is the same as the sum of the squares on the four sides of any parallelogram. Illustrate this conclusion with the following examples:

(a) $\mathbf{x} = (1, 2), \mathbf{y} = (2, 1)$,
(b) $\mathbf{x} = (-1, 0), \mathbf{y} = (1, 1)$,
(c) $\mathbf{x} = (a, 0), \mathbf{y} = (0, b)$, where $a \neq 0$ and $b \neq 0$.

Exercise 11   Illustrate the Pythagorean theorem, namely (1.3.12), with the following pairs of vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^2$ after verifying that $\mathbf{x}, \mathbf{y}$ are perpendicular to each other (that is, $\langle \mathbf{x}, \mathbf{y} \rangle = 0$):

(a) $\mathbf{x} = (1, -1), \mathbf{y} = (2, 2)$,
(b) $\mathbf{x} = (4, 1), \mathbf{y} = (-2, 8)$,
(c) $\mathbf{x} = (a, 0), \mathbf{y} = (0, b)$ with $a \neq 0$ and $b \neq 0$, as in (c) of Exercise 10.

Exercise 12   Illustrate the Pythagorean theorem, namely (1.3.12), with the following pairs of functions $f, g \in \widetilde{L}_2[0, 1]$ after verifying that $\langle f, g \rangle = 0$:

(a) $f(x) = \sin 2\pi x, g(x) = 1$,
(b) $f(x) = \sin 2\pi x, g(x) = \cos 2\pi x$,
(c) $f(x) = \cos 2\pi j x, g(x) = \cos 2\pi k x$, where j and k are distinct integers.

Exercise 13   Show that if $\mathbb{W}$ is a subspace of an inner-product space $\mathbb{V}$, then $\mathbb{W}^{\perp}$, as defined in (1.3.11), is also a subspace of $\mathbb{V}$.

Exercise 14   As a continuation of Exercise 13, show that $\mathbb{W} \cap \mathbb{W}^{\perp} = \{\mathbf{0}\}$.

Exercise 15   Let $\{\mathbf{x}_j\}$ be a set of vectors in a vector space $\mathbb{V}$. Show that span$\{\mathbf{x}_j\}$ is a subspace of $\mathbb{V}$.

Exercise 16   Let $\mathbf{x} = (1, -1, 1) \in \mathbb{R}^3$ and $\mathbb{W} = $ span$\{\mathbf{x}\}$. What is the best you can say about $\mathbb{W}^{\perp}$?

Exercise 17   Let $\mathbf{x}_1 = (1, -1, 3, -1), \mathbf{x}_2 = (1, 2, 2, 1) \in \mathbb{R}^4$ and denote $\mathbb{W} = $ span$\{\mathbf{x}_1, \mathbf{x}_2\}$. What is the best you can say about $\mathbb{W}^{\perp}$?

1.4 Bases of Sequence and Function Spaces

The concepts of (algebraic) linear dependency and independency, bases, orthogonality, orthogonal projection and Gram-Schmidt process from elementary Linear Algebra are reviewed in this section. In addition, these concepts will be extended to infinite-dimensional vector spaces such as ℓ_p and $\widetilde{L}_p(J)$.

Let $\mathbb{V}$ be a vector space over some scalar field $\mathbb{F}$. Recall that a finite set $S = \{v_k : 1 \le k \le n\} \subset \mathbb{V}$ is said to be **linearly dependent**, if there exist $c_1, \ldots, c_n \in \mathbb{F}$, not all zero, such that

$$\sum_{k=1}^{n} c_k v_k = 0.$$

If S is not linearly dependent, then it is said to be **linearly independent**. In other words, S is linearly independent if and only if the only solution of the equation

$$\sum_{k=1}^{n} c_k v_k = 0$$

for $c_1, \ldots, c_n$ in $\mathbb{F}$ is the zero solution; that is, $c_1 = \cdots = c_n = 0$. Hence, if the zero vector 0 is in S, then S is a linearly dependent set of vectors.

Example 1 Consider the vectors $x_1 = (1, 0, 2)$, $x_2 = (0, 1, 1)$, $x_3 = (-1, 1, 0)$ in the vector space $\mathbb{R}^3$. Determine whether or not the set $\{x_1, x_2, x_3\}$ is linearly independent.

Solution Consider the equation

$$c_1 x_1 + c_2 x_2 + c_3 x_3 = 0,$$

or

$$c_1(1, 0, 2) + c_2(0, 1, 1) + c_3(-1, 1, 0) = (0, 0, 0);$$

that is,

$$\begin{cases} c_1 - c_3 = 0, \\ c_2 + c_3 = 0, \\ 2c_1 + c_2 = 0. \end{cases} \tag{1.4.1}$$

Putting $c_1 = c_3$ from the first equation to the third equation yields

$$\begin{cases} c_2 + c_3 = 0, \\ c_2 + 2c_3 = 0, \end{cases}$$

so that $c_3 = 2c_3$, or $c_3 = 0 \Rightarrow c_2 = 0 \Rightarrow c_1 = 0$. Hence, $c_1 = c_2 = c_3 = 0$, and the set $\{x_1, x_2, x_3\}$ is linearly independent. ∎

Definition 1 | **Basis for finite-dimensional space** | *A finite set $S = \{v_k : 1 \le k \le n\}$ of vectors in a vector space $\mathbb{V}$ is said to be a basis of $\mathbb{V}$ if S is linearly independent and any $x \in \mathbb{V}$ can be expressed as*

$$x = \sum_{k=1}^{n} c_k v_k$$

for some $c_k \in \mathbb{F}$. In this case, $\mathbb{V}$ is said to be a finite-dimensional space, and n is called the dimension of $\mathbb{V}$.

For example, $\mathbb{R}^n, \mathbb{C}^n$ and $\mathbb{R}^{m,n}, \mathbb{C}^{m,n}$ are finite-dimensional vector spaces. In particular, the set of

$$\mathbf{e}_1 = (1, 0, 0, \ldots, 0), \mathbf{e}_2 = (0, 1, 0, \ldots, 0), \ldots, \mathbf{e}_n = (0, 0, \ldots, 0, 1) \quad (1.4.2)$$

is a basis for $\mathbb{R}^n$ and $\mathbb{C}^n$. We leave the proof as an exercise (see Exercise 4). There is rich theory on the linear independence and bases for finite-dimensional spaces $\mathbb{V}$ in elementary Linear Algebra. For example, it can be shown that the dimension n of $\mathbb{V}$ is independent of the choices of bases; and for $\mathbb{V} = \mathbb{R}^n$, $S = \{\mathbf{v}_k : 1 \leq k \leq n\}$ is a basis for $\mathbb{R}^n$ if and only if S is linearly independent, which is equivalent to that the $n \times n$ matrix $A = [\mathbf{v}_1 \mathbf{v}_2 \cdots \mathbf{v}_n]$, with $\mathbf{v}_1, \ldots, \mathbf{v}_n$ as columns of A, is nonsingular.

A vector space is called an **infinite-dimensional vector space** if it is not finite-dimensional. To extend the notion of basis to an infinite-dimensional vector space, let us first give the following equivalent definition of basis for a finite-dimensional vector space $\mathbb{V}$.

Remark 1 A set $S = \{\mathbf{v}_k : 1 \leq k \leq n\}$ in a finite-dimensional vector space $\mathbb{V}$ is a basis of $\mathbb{V}$ if and only if each $\mathbf{x} \in \mathbb{V}$ can be written uniquely as a linear combination of $\mathbf{v}_k$, $1 \leq k \leq n$.

Indeed, if S is a basis of $\mathbb{V}$, then for any $\mathbf{x} \in \mathbb{V}$, writing

$$\mathbf{x} = \sum_{k=1}^{n} c_k \mathbf{v}_k$$

as well as

$$\mathbf{x} = \sum_{k=1}^{n} d_k \mathbf{v}_k,$$

and taking the difference of these two linear combinations, we have

$$\sum_{k=1}^{n} (c_k - d_k) \mathbf{v}_k = \mathbf{0}.$$

Hence, the linear independence of $\{\mathbf{v}_k : 1 \leq k \leq n\}$ implies that $c_k = d_k, k = 1, \ldots, n$. That is, the linear combination representation is unique.

Conversely, suppose each $\mathbf{x} \in \mathbb{V}$ can be expressed uniquely as a linear combination of $\mathbf{v}_k$, $1 \leq k \leq n$. Then in the first place, the algebraic span of S is all of $\mathbb{V}$. In addition, suppose $\sum_{k=1}^{n} c_k \mathbf{v}_k = \mathbf{0}$. Then this is a linear combination representation of the zero

vector $\mathbf{0}$, which is a vector in $\mathbb{V}$. But since $\mathbf{0} = \sum_{k=1}^{n} 0\mathbf{v}_k$, it follows from the unique linear combination representation of the zero vector $\mathbf{0}$ that $c_k = 0, k = 1, \ldots, n$. That is, the set $\{\mathbf{v}_k : 1 \le k \le n\}$ is linearly independent, and is therefore a basis of $\mathbb{V}$. ∎

Motivated by the above remark, we introduce the notion of (algebraic) basis for an infinite-dimensional vector space as follows.

Definition 2 │ **Algebraic basis for infinite-dimensional space** │   *An infinite set $S = \{\mathbf{v}_k : k = 1, 2, \ldots\}$ of vectors in an infinite-dimensional vector space $\mathbb{V}$ is said to be an algebraic basis for $\mathbb{V}$, if each $\mathbf{x} \in \mathbb{V}$ can be expressed uniquely as a finite linear combination of S; that is, each $\mathbf{x} \in \mathbb{V}$ is written uniquely as*

$$\mathbf{x} = \sum_{j=1}^{J} c_j \mathbf{v}_{k_j}$$

for some integer $J \ge 1$, constants $c_j \in \mathbb{F}$, and vectors $\mathbf{v}_{k_j} \in S$ that depend on $\mathbf{x}$.

To give an example, let us first introduce the "Kronecker delta" symbol, defined by

$$\delta_j = \begin{cases} 1, & \text{if } j = 0, \\ 0, & \text{if } j \ne 0, \end{cases} \tag{1.4.3}$$

where j is any integer. In addition, for each fixed integer $k \in \mathbb{Z}$, the delta sequence δ_k is defined by

$$\delta_k = \{\delta_{k-j}\}_{j=\cdots,-1,0,1,\ldots};$$

that is, $\delta_k = (\ldots, 0, \ldots, 0, 1, 0, \ldots)$, where the value 1 occurs at the kth entry, while all the other entries are equal to zero. It is clear that each delta sequence δ_k is in ℓ_p for any p with $0 \le p \le \infty$. In the following, we will see that while $S = \{\delta_k : k = 0, \pm 1, \pm 2, \ldots\}$ is an algebraic basis of ℓ_0, it is not an algebraic basis of ℓ_p for $0 < p \le \infty$.

Example 2 Verify that $S = \{\delta_k : k = 0, \pm 1, \pm 2, \cdots\}$ is an algebraic basis for ℓ_0.

Solution   Indeed, for any $\mathbf{x} \in \ell_0$, there are only finitely many entries x_k of $\mathbf{x}$ that are nonzero. Thus, we have $x_k = 0$ for all $|k| > K$ for some positive integer K. Therefore $\mathbf{x}$ can be written as a finite linear combination of S, namely:

$$\mathbf{x} = (\ldots, 0, x_{-K}, \ldots, x_K, 0, \ldots) = \sum_{j=-K}^{K} x_k \delta_k.$$

To show that the above expression is unique, let $\mathbf{x} = \sum_{j=-K}^{K} d_k \boldsymbol{\delta}_k$, and take the difference of these two sums to arrive at

$$\sum_{j=-K}^{K} (x_k - d_k)\boldsymbol{\delta}_k = \mathbf{x} - \mathbf{x} = \mathbf{0}.$$

That is,

$$(\ldots, 0, x_{-K} - d_{-K}, \ldots, x_K - d_K, 0, \ldots) = \mathbf{0} = (\ldots, 0, 0, \ldots, 0, 0, \ldots),$$

so that $x_k = d_k$, $-K \le k \le K$. Therefore, every $\mathbf{x} \in \ell_0$ can be written uniquely as a finite linear combination of $S = \{\boldsymbol{\delta}_k \,:\, k = 0, \pm 1, \pm 2, \ldots\}$. That is, S is an algebraic basis of ℓ_0.   ∎

Example 3   Give an example to illustrate that $S = \{\boldsymbol{\delta}_k : k = 0, \pm 1, \pm 2, \cdots\}$ is not an algebraic basis of ℓ_p for $0 < p \le \infty$.

Solution   For each real number p with $0 < p < \infty$, the sequence $\mathbf{y}_p = \{y_k \,:\, -\infty < k < \infty\}$, where $y_k = (1 + k^2)^{-1/p}$, is in ℓ_p, since

$$\sum_{k=-\infty}^{\infty} |y_k|^p = \sum_{k=-\infty}^{\infty} \frac{1}{1 + k^2} < \infty.$$

However, since each term y_k of the sequence is non-zero, it is clear that $\mathbf{y}_p$ cannot be written as a finite linear combination of $S = \{\boldsymbol{\delta}_k : k = 0, \pm 1, \pm 2, \ldots\}$.

For $p = \infty$, consider $\mathbf{y}_\infty = (\ldots 1, 1, \ldots, 1 \ldots)$, where each term of the sequence $\mathbf{y}_\infty$ is equal to 1. Hence, $\mathbf{y}_\infty \in \ell_\infty$ but cannot be written as a finite linear combination of $S = \{\boldsymbol{\delta}_k : k = 0, \pm 1, \pm 2, \ldots\}$ either.   ∎

The notion of "algebraic bases", as introduced in Definition 2, is limited only to "finite" linear combinations. To extend to "infinite linear combinations" which are actually infinite series, it is clear that the concept of convergence must be taken into consideration. In other words, the notion of "measurement" of vectors in $\mathbb{V}$ is again needed. For convenience, we will only consider the norm measurement induced by some inner product. Therefore, in what follows in this section, the vector space $\mathbb{V}$ to be studied is always an inner-product space over some scalar field $\mathbb{F} = \mathbb{C}$ or $\mathbb{R}$ with inner product $\langle \cdot, \cdot \rangle$ and norm $|| \cdot || = \sqrt{\langle \cdot, \cdot \rangle}$.

Let $\mathbf{v}_k, k = 1, 2, \ldots$, be vectors in $\mathbb{V}$. We say that the infinite series $\sum_{k=1}^{\infty} c_k \mathbf{v}_k$, where $c_k \in \mathbb{F}$, is convergent in $\mathbb{V}$, if the sequence of its partial sums converges to some vector $\mathbf{x} \in \mathbb{V}$; that is,

$$\lim_{n \to \infty} \left\| \mathbf{x} - \sum_{k=1}^{n} c_k \mathbf{v}_k \right\| = 0. \tag{1.4.4}$$

In this case, $\mathbf{x}$ is called the limit of the series, and we write

$$\mathbf{x} = \sum_{k=1}^{\infty} c_k \mathbf{v}_k. \tag{1.4.5}$$

Of course, it should be obvious that the one-sided infinite sum (or infinite series) in the above formulation can be easily extended to the two-sided infinite sum without further discussion, but only with minimum change of notation, such as

$$\sum_{k=-\infty}^{\infty} c_k \mathbf{v}_k,$$

if the set $\{\mathbf{v}_k : k = 1, 2, \ldots\}$ is replaced by $\{\mathbf{v}_k : k = 0, \pm 1, \pm 2, \ldots\}$. We will encounter such situations quite frequently, particularly in Chapters 6-10 of this book.

The following definitions are valid for all normed (linear) spaces to be introduced in Sect. 1.5, but we will only apply it to the norm induced by some inner product in this section.

Definition 3 | **Complete set and basis of infinite-dimensional space** | *Let $\mathbb{V}$ be a normed space over some scalar field $\mathbb{F}$ with norm $\|\cdot\|$. A set $S = \{\mathbf{v}_k, k = 1, 2, \cdots\}$ of vectors in $\mathbb{V}$ is said to be complete if every $\mathbf{x} \in \mathbb{V}$ can be represented as in (1.4.5) for some $c_k \in \mathbb{F}$. If in addition, the representation is unique, then S is called a basis for $\mathbb{V}$. These definitions apply to $\{\mathbf{v}_k : k = 0, \pm 1, \pm 2, \cdots\}$, with the one-sided infinite sum in (1.4.5) replaced by the two-sided infinite sum.*

In the above definition of bases, the uniqueness of the infinite series representation for $\mathbf{x} \in \mathbb{V}$ means that the coefficients $c_k, k = 1, 2, \ldots$ are uniquely determined by $\mathbf{x}$. This notion is equivalent to that of linear independence of an infinite set of vectors, introduced in the following definition.

Definition 4 | **Linear independence of infinite set** | *A set $S = \{\mathbf{v}_k : k = 1, 2, \ldots\}$ of vectors in a normed space $\mathbb{V}$ over a scalar field $\mathbb{F}$ is said to be linearly dependent if there exist $c_k \in \mathbb{F}, k = 1, 2, \ldots$, not all equal to zero, such that*

$$\sum_{k=1}^{\infty} c_k \mathbf{v}_k = \mathbf{0}.$$

A set of vectors in $\mathbb{V}$ is linearly independent if it is not linearly dependent.

In other words, a set $\{\mathbf{v}_k : k = 1, 2, \ldots\}$ of vectors in $\mathbb{V}$ is linearly independent, if $\sum_{k=1}^{\infty} c_k \mathbf{v}_k = \mathbf{0}$ (meaning that the series converges to $\mathbf{0}$) for some $c_k \in \mathbb{F}, k = 1, 2, \ldots$, then c_k must be zero, for all k.

In view of the above definition of linear independence for an infinite set of vectors in a normed space $\mathbb{V}$, an equivalent definition of basis for $\mathbb{V}$ can be stated as follows:

Remark 2 | **Linear independence + completeness = basis** | A set $S = \{\mathbf{v}_k : k = 1, 2, \ldots\}$ of vectors in a normed space $\mathbb{V}$ is a basis of $\mathbb{V}$ if S is linearly independent and complete in $\mathbb{V}$.   ∎

For an inner-product space $\mathbb{V}$, the notion of orthogonal basis for a finite-dimensional space in an elementary course of Linear Algebra can also be extended to infinite-dimensional inner-product spaces, and in fact, to the notion of a pair of "dual bases", as follows:

Definition 5 | **Orthogonal and dual bases** | *Let $\mathbb{V}$ be an inner-product space over some scalar field $\mathbb{F}$ with inner product $\langle \cdot, \cdot \rangle$ and norm $|| \cdot || = \sqrt{\langle \cdot, \cdot \rangle}$.*

(a) *A basis $S = \{\mathbf{v}_k : k = 1, 2, \ldots\}$ of $\mathbb{V}$ is called an orthogonal basis of $\mathbb{V}$ if*

$$\langle \mathbf{v}_k, \mathbf{v}_j \rangle = 0, \ j \neq k, \ j, k = 1, 2, \ldots .$$

If, in addition, $\|\mathbf{v}_k\| = 1$ for all k; that is,

$$\langle \mathbf{v}_k, \mathbf{v}_j \rangle = \delta_{k-j}, \ j, k = 1, 2, \ldots ,$$

then S is called an orthonormal basis of $\mathbb{V}$.
(b) *Two bases $\{\mathbf{v}_k : k = 1, 2, \ldots\}$ and $\{\widetilde{\mathbf{v}}_k : k = 1, 2, \ldots\}$ of $\mathbb{V}$ are said to be dual bases of $\mathbb{V}$, if*

$$\langle \mathbf{v}_k, \widetilde{\mathbf{v}}_j \rangle = \delta_{k-j}, \ j, k = 1, 2, \ldots .$$

Remark 3 | **Orthogonality + Completeness $\Leftrightarrow$ Orthogonal basis** |
It is not difficult to show that if a set $S = \{\mathbf{v}_k : k = 1, 2, \ldots\}$ of nonzero vectors $\mathbf{v}_k$ in an inner-product space $\mathbb{V}$ is orthogonal, then S is linearly independent (see Exercise 9), and hence, S is an orthogonal basis for $\mathbb{V}$ if and only if S is complete. ∎

Example 4 Identify whether $S = \{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ and $\widetilde{S} = \{\widetilde{\mathbf{v}}_1, \widetilde{\mathbf{v}}_2, \widetilde{\mathbf{v}}_3\}$ of $\mathbb{R}^3$ are dual bases or not, where

$$\mathbf{v}_1 = (2, 1, -1), \mathbf{v}_2 = (1, 0, -1), \mathbf{v}_3 = (1, 1, 1),$$
$$\widetilde{\mathbf{v}}_1 = (-1, 2, -1), \widetilde{\mathbf{v}}_2 = (2, -3, 1), \widetilde{\mathbf{v}}_3 = (1, -1, 1).$$

Solution It is easy to verify that both S and $\widetilde{S}$ are bases of $\mathbb{R}^3$. By simple computation, we have

$$\langle \mathbf{v}_1, \widetilde{\mathbf{v}}_1 \rangle = 2(-1) + 1(2) + (-1)(-1) = 1,$$
$$\langle \mathbf{v}_1, \widetilde{\mathbf{v}}_2 \rangle = 2(2) + 1(-3) + (-1)(1) = 0,$$
$$\langle \mathbf{v}_1, \widetilde{\mathbf{v}}_3 \rangle = 2(1) + 1(-1) + (-1)(1) = 0.$$

Similarly, we can obtain $\langle \mathbf{v}_k, \widetilde{\mathbf{v}}_j \rangle = \delta_{k-j}$ for $k = 2, 3, 1 \leq j \leq 3$. Thus, S and $\widetilde{S}$ are dual bases of $\mathbb{R}^3$. ∎

Example 5 Verify that the set $S = \{\delta_k : k = 0, \pm 1, \pm 2, \ldots\}$ as discussed in Example 2 is an orthonormal basis of ℓ_2.

Solution Let $\mathbf{x} = \{x_k\}$ be a sequence in ℓ_2. It follows from the definition of δ_k that

$$\mathbf{x} = \sum_{k=-\infty}^{\infty} x_k \delta_k,$$

which converges in the sense of (1.4.4), since

$$\left\| \mathbf{x} - \sum_{|k| \leq N} x_k \delta_k \right\|_2^2 = \sum_{|k| > N} |x_k|^2 \to 0$$

(see Exercise 5). Clearly this representation for $\mathbf{x}$ is unique, so that S is a basis for ℓ_2.

In addition, for $j, k = 0, \pm 1, \pm 2, \ldots$,

$$\langle \delta_k, \delta_j \rangle = \sum_{m=-\infty}^{\infty} \delta_{k-m} \delta_{j-m} = \delta_{j-k},$$

and hence, S is an orthonormal basis for ℓ_2. ∎

Next, we introduce the concept of orthogonal projection. Let $\mathbf{w}_1, \ldots, \mathbf{w}_n$ be nonzero orthogonal vectors in an inner-product ssspace $\mathbb{V}$ over some scalar field $\mathbb{F}$, with inner product $\langle \cdot, \cdot \rangle$ and norm $\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$, $\mathbf{x} \in \mathbb{V}$, and consider the algebraic span

$$\mathbb{W} = \text{span}\{\mathbf{w}_1, \ldots, \mathbf{w}_n\}.$$

For any $\mathbf{v} \in \mathbb{V}$, the vector $P\mathbf{v} \in \mathbb{W}$, defined by

$$P\mathbf{v} = c_1(\mathbf{v})\mathbf{w}_1 + c_2(\mathbf{v})\mathbf{w}_2 + \cdots + c_n(\mathbf{v})\mathbf{w}_n, \tag{1.4.6}$$

with

$$c_j(\mathbf{v}) = \frac{\langle \mathbf{v}, \mathbf{w}_j \rangle}{\|\mathbf{w}_j\|^2}, \quad j = 1, 2, \ldots, n, \tag{1.4.7}$$

is called the **orthogonal projection** of $\mathbf{v}$ onto $\mathbb{W}$. Let $\mathbb{W}^{\perp}$ be the orthogonal complement of $\mathbb{W}$ in $\mathbb{V}$ defined by (1.3.11) on p. 29.

Theorem 1 | **Properties of orthogonal projection** | *Let $P\mathbf{v}$ be the orthogonal projection of $\mathbf{v}$ onto $\mathbb{W}$ as defined by (1.4.6). Then*

(a) $P\mathbf{v} = \mathbf{v}$ *if $\mathbf{v} \in \mathbb{W}$.*
(b) $P\mathbf{v} = \mathbf{0}$ *if $\mathbf{v} \in \mathbb{W}^{\perp}$.*

(c) $\mathbf{v} - P\mathbf{v} \in \mathbb{W}^{\perp}$; *that is*, $\langle \mathbf{v} - P\mathbf{v}, \mathbf{w} \rangle = 0$ *for all* $\mathbf{w} \in \mathbb{W}$.

Since (a) and (b) are at least intuitively clear, it is safe to leave the proof as an exercise. To prove (c), it is sufficient to verify that $\langle \mathbf{v} - P\mathbf{v}, \mathbf{w}_k \rangle = 0$ for $1 \le k \le n$. This indeed holds, since

$$
\begin{aligned}
\langle \mathbf{v} - P\mathbf{v}, \mathbf{w}_k \rangle &= \langle \mathbf{v}, \mathbf{w}_k \rangle - \langle P\mathbf{v}, \mathbf{w}_k \rangle \\
&= \langle \mathbf{v}, \mathbf{w}_k \rangle - \sum_{j=1}^{n} c_j(\mathbf{v}) \langle \mathbf{w}_j, \mathbf{w}_k \rangle \\
&= \langle \mathbf{v}, \mathbf{w}_k \rangle - c_k(\mathbf{v}) \langle \mathbf{w}_k, \mathbf{w}_k \rangle \\
&= \langle \mathbf{v}, \mathbf{w}_k \rangle - \frac{\langle \mathbf{v}, \mathbf{w}_k \rangle}{\|\mathbf{w}_k\|^2} \langle \mathbf{w}_k, \mathbf{w}_k \rangle = 0. \qquad \blacksquare
\end{aligned}
$$

Example 6 Let $\mathbf{w}_1 = (0, 4, 2)$, $\mathbf{w}_2 = (-5, -1, 2)$, $\mathbb{W} = \mathrm{span}\{\mathbf{w}_1, \mathbf{w}_2\}$. For the given vector $\mathbf{v} = (-5, 4, 12)$, compute the orthogonal projection $P\mathbf{v}$ of $\mathbf{v}$ onto $\mathbb{W}$. Repeat this for $\mathbf{u} = (-10, 8, 9)$.

Solution We must first verify that the two vectors $\mathbf{w}_1$ and $\mathbf{w}_2$ are orthogonal to each other, so that (1.4.6) and (1.4.7) can be applied to compute $P\mathbf{v}$, $P\mathbf{u}$. In this regard, we remark that if they are not orthogonal, then the Gram-Schmidt process to be discussed later in this section can be followed to replace $\mathbf{w}_2$ by another non-zero vector in $\mathrm{span}\{\mathbf{w}_1, \mathbf{w}_2\}$, which is orthogonal to $\mathbf{w}_1$, before carrying out the computations below. For these two vectors $\mathbf{w}_1$ and $\mathbf{w}_2$ in this example, it is clear they are orthogonal.

By applying (1.4.6) and (1.4.7), we have

$$
\begin{aligned}
P\mathbf{v} &= \frac{\langle \mathbf{v}, \mathbf{w}_1 \rangle}{\|\mathbf{w}_1\|^2} \mathbf{w}_1 + \frac{\langle \mathbf{v}, \mathbf{w}_2 \rangle}{\|\mathbf{w}_2\|^2} \mathbf{w}_2 \\
&= \frac{0 + 16 + 24}{20} \mathbf{w}_1 + \frac{25 - 4 + 24}{30} \mathbf{w}_2 \\
&= 2(0, 4, 2) + \frac{3}{2}(-5, -1, 2) = \left(-\frac{15}{2}, \frac{13}{2}, 7 \right).
\end{aligned}
$$

Similarly, we have, for $\mathbf{u} = (-10, 8, 9)$,

$$
\begin{aligned}
P\mathbf{u} &= \frac{\langle \mathbf{u}, \mathbf{w}_1 \rangle}{\|\mathbf{w}_1\|^2} \mathbf{w}_1 + \frac{\langle \mathbf{u}, \mathbf{w}_2 \rangle}{\|\mathbf{w}_2\|^2} \mathbf{w}_2 \\
&= \frac{0 + 32 + 18}{20} \mathbf{w}_1 + \frac{50 - 8 + 18}{30} \mathbf{w}_2 \\
&= \frac{5}{2}(0, 4, 2) + 2(-5, -1, 2) = (-10, 8, 9).
\end{aligned}
$$

Observe from the above computation that $P\mathbf{u} = \mathbf{u} \in \mathbb{W}$. Hence, we may conclude that $\mathrm{dist}(\mathbf{u}, \mathbb{W}) = 0$ without any work. Note that $P\mathbf{u}$ is a linear combination of $\mathbf{w}_1$

and $\mathbf{w}_2$, from the above computation, namely: $\mathbf{u} = \dfrac{5}{2}\mathbf{w}_1 + 2\mathbf{w}_2$. This conclusion is also assured by statement (a) of Theorem 1.

To illustrate $P\mathbf{v}$ geometrically, the vectors $\mathbf{v}$, $\mathbf{w}_1$ and $\mathbf{w}_2$ are in the Euclidean space $\mathbb{V} = \mathbb{R}^3$, and the subspace $\mathbb{W}$ is a plane in $\mathbb{R}^3$ that contains the two vectors $\mathbf{w}_1$ and $\mathbf{w}_2$ (see Fig. 1.3). The projection $P\mathbf{v}$ of $\mathbf{v}$ onto $\mathbb{W}$ is the vector $(-\dfrac{15}{2}, \dfrac{13}{2}, 7)$ on $\mathbb{W}$ and

$$\mathbf{v} - P\mathbf{v} = (-5, 4, 12) - \left(-\frac{15}{2}, \frac{13}{2}, 7\right) = \frac{5}{2}(1, -1, 2)$$

is the vector orthogonal to $\mathbb{W}$. ∎

Example 7 Let the functions g_1, g_2, g_3 be defined as follows:

$$g_1(x) = \frac{1}{\sqrt{\pi}}, \quad g_2(x) = \frac{\sqrt{2}}{\sqrt{\pi}}\cos x, \quad g_3(x) = \frac{\sqrt{2}}{\sqrt{\pi}}\cos 2x.$$

These functions are in $\widetilde{L}_2[0, \pi]$ and constitute an orthonormal basis of its span $\mathbb{W} = \text{span}\{g_1, g_2, g_3\}$ (see Exercise 6). Compute the orthogonal projection Pf of the function $f(x) = x^2$ onto $\mathbb{W}$.

Solution It is easy to verify that the functions g_j, $1 \le j \le 3$ are orthonormal to one another and that they have unit length, namely: $\|g_j\| = \|g_j\|_2 = 1$ for $j = 1, 2, 3$. Therefore, by the definition of Pf, we have
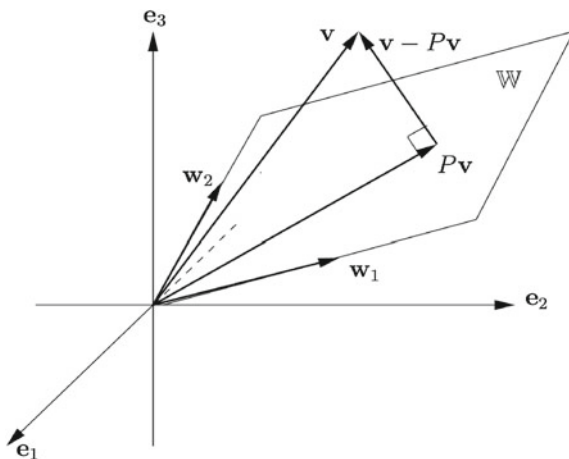

Fig. 1.3 Orthogonal projection $P\mathbf{v}$ in $\mathbb{W}$

$$
\begin{aligned}
Pf(x) &= \frac{\langle f, g_1 \rangle}{\|g_1\|^2} g_1(x) + \frac{\langle f, g_2 \rangle}{\|g_2\|^2} g_2(x) + \frac{\langle f, g_3 \rangle}{\|g_3\|^2} g_3(x) \\
&= \int_0^\pi x^2 \frac{1}{\sqrt{\pi}} dx \, g_1(x) \\
&\quad + \int_0^\pi x^2 \frac{\sqrt{2}}{\sqrt{\pi}} \cos x \, dx \, g_2(x) + \int_0^\pi x^2 \frac{\sqrt{2}}{\sqrt{\pi}} \cos 2x \, dx \, g_3(x) \\
&= \frac{1}{\sqrt{\pi}} \frac{\pi^3}{3} g_1(x) + \frac{\sqrt{2}}{\sqrt{\pi}} (-2\pi) g_2(x) + \frac{\sqrt{2}}{\sqrt{\pi}} \frac{\pi}{2} g_3(x) \\
&= \frac{\pi^2}{3} - 4\cos x + \cos 2x. \qquad \blacksquare
\end{aligned}
$$

Next we consider the problem of best approximation of any vector $\mathbf{v} \in \mathbb{V}$ by vectors from $\mathbb{W}$; namely, for any given $\mathbf{v} \in \mathbb{V}$, we wish to find some $\mathbf{w}_0 \in \mathbb{W}$, such that $\mathbf{w}_0$ is the closest to the given $\mathbf{v}$ among all $\mathbf{w} \in \mathbb{W}$. The following result solves this problem by assuring that the orthogonal projection $P\mathbf{v}$ is the desired $\mathbf{w}_0 \in \mathbb{W}$.

Theorem 2 $\boxed{\text{Orthogonal projection yields best approximation}}$ *Let* $\mathbb{W}$ *be any finite-dimensional subspace of an inner-product space* $\mathbb{V}$ *and* $\mathbf{v}$ *any vector in* $\mathbb{V}$. *Then the projection* $P\mathbf{v}$ *of* $\mathbf{v}$ *onto* $\mathbb{W}$ *provides the best approximation of* $\mathbf{v}$ *from the subspace* $\mathbb{W}$. *Precisely,*

$$
\|\mathbf{v} - P\mathbf{v}\| = \min_{\mathbf{w} \in \mathbb{W}} \|\mathbf{v} - \mathbf{w}\|.
$$

Proof To prove this result, let $\mathbf{w}_1, \ldots, \mathbf{w}_n$ be any orthogonal basis of $\mathbb{W}$. Later in this section, we will derive a procedure, called the Gram-Schmidt process, for computing an orthonormal basis of $\mathbb{W}$ from any given basis of $\mathbb{W}$.

Let $\mathbf{w} \in \mathbb{W}$ be arbitrarily chosen. Then by (c) of Theorem 1, since $P\mathbf{v} - \mathbf{w} \in \mathbb{W}$, the vector $\mathbf{v} - P\mathbf{v}$ is orthogonal to $P\mathbf{v} - \mathbf{w}$. Thus, it follows from the Pythagorean theorem (see Theorem 4 in Sect. 1.3 on p.24) that

$$
\begin{aligned}
\|\mathbf{v} - \mathbf{w}\|^2 &= \|(\mathbf{v} - P\mathbf{v}) + (P\mathbf{v} - \mathbf{w})\|^2 \\
&= \|\mathbf{v} - P\mathbf{v}\|^2 + \|P\mathbf{v} - \mathbf{w}\|^2,
\end{aligned}
$$

so that

$$
\|\mathbf{v} - P\mathbf{v}\|^2 \leq \|\mathbf{v} - \mathbf{w}\|^2
$$

for all $\mathbf{w} \in \mathbb{W}$. $\blacksquare$

The non-negative real number $\min_{\mathbf{w} \in \mathbb{W}} \|\mathbf{v} - \mathbf{w}\|$ is called the distance of $\mathbf{v}$ from $\mathbb{W}$, denoted by $\text{dist}(\mathbf{v}, \mathbb{W})$. The above theorem assures that this distance is attained by the projection $P\mathbf{v}$. To derive an explicit formula of $\text{dist}(\mathbf{v}, \mathbb{W})$, we rely on the orthogonal basis $\mathbf{w}_1, \ldots, \mathbf{w}_n$, namely: its square is given by

$$\|\mathbf{v} - P\mathbf{v}\|^2 = \left\langle \mathbf{v} - \sum_{j=1}^{n} c_j(\mathbf{v})\mathbf{w}_j, \mathbf{v} - \sum_{j=1}^{n} c_j(\mathbf{v})\mathbf{w}_j \right\rangle$$

$$= \langle \mathbf{v}, \mathbf{v} \rangle - \sum_{j=1}^{n} \overline{c_j(\mathbf{v})} \langle \mathbf{v}, \mathbf{w}_j \rangle - \sum_{j=1}^{n} c_j(\mathbf{v}) \langle \mathbf{w}_j, \mathbf{v} \rangle + \sum_{j=1}^{n} \|c_j(\mathbf{v})\|^2 \langle \mathbf{w}_j, \mathbf{w}_j \rangle$$

$$= \|\mathbf{v}\|^2 - \sum_{j=1}^{n} \frac{|\langle \mathbf{v}, \mathbf{w}_j \rangle|^2}{\|\mathbf{w}_j\|^2} - \sum_{j=1}^{n} \frac{|\langle \mathbf{v}, \mathbf{w}_j \rangle|^2}{\|\mathbf{w}_j\|^2} + \sum_{j=1}^{n} \frac{|\langle \mathbf{v}, \mathbf{w}_j \rangle|^2}{\|\mathbf{w}_j\|^2}$$

$$= \|\mathbf{v}\|^2 - \sum_{j=1}^{n} \frac{|\langle \mathbf{v}, \mathbf{w}_j \rangle|^2}{\|\mathbf{w}_j\|^2}. \tag{1.4.8}$$

Let us summarize this result in the following theorem.

Theorem 3 $\boxed{\text{Bessel's inequality}}$ *Let* $\mathbb{W}$ *be any subspace of an inner-product space* $\mathbb{V}$ *with orthogonal basis* $\mathbf{w}_1, \ldots, \mathbf{w}_n$. *Then for any* $\mathbf{v} \in \mathbb{V}$,

(a)

$$\|\mathbf{v} - P\mathbf{v}\|^2 = \left(\text{dist}(\mathbf{v}, \mathbb{W})\right)^2 = \|\mathbf{v}\|^2 - \sum_{j=1}^{n} \frac{|\langle \mathbf{v}, \mathbf{w}_j \rangle|^2}{\|\mathbf{w}_j\|^2}. \tag{1.4.9}$$

(b) *Bessel's inequality:*

$$\sum_{j=1}^{n} \frac{|\langle \mathbf{v}, \mathbf{w}_j \rangle|^2}{\|\mathbf{w}_j\|^2} \leq \|\mathbf{v}\|^2. \tag{1.4.10}$$

It is clear that Bessel's inequality (1.4.10) is a trivial consequence of (1.4.9), since $\text{dist}(\mathbf{v}, \mathbb{W}) \geq 0$.

Example 8 Let $\mathbf{v}$ and $\mathbb{W}$ be the vector and space considered in Example 6. Find the distance $\text{dist}(\mathbf{v}, \mathbb{W})$ of $\mathbf{v}$ to $\mathbb{W}$.

Solution By applying (1.4.9), we have

$$\left(\text{dist}(\mathbf{v}, \mathbb{W})\right)^2 = \|\mathbf{v}\|^2 - \frac{|\langle \mathbf{v}, \mathbf{w}_1 \rangle|^2}{\|\mathbf{w}_1\|^2} - \frac{|\langle \mathbf{v}, \mathbf{w}_2 \rangle|^2}{\|\mathbf{w}_2\|^2}$$

$$= 185 - \frac{40^2}{20} - \frac{45^2}{30} = \frac{75}{2},$$

so that $\text{dist}(\mathbf{v}, \mathbb{W}) = \dfrac{5\sqrt{6}}{2}$. Alternatively, we may compute $\text{dist}(\mathbf{v}, \mathbb{W})$ by observing it is the same as $\|\mathbf{v} - \overset{\circ}{P}\mathbf{v}\|$, namely:

$$\text{dist}(\mathbf{v}, \mathbb{W}) = \|\mathbf{v} - P\mathbf{v}\| = \left\|(-5, 4, 12) - \left(-\frac{15}{2}, \frac{13}{2}, 7\right)\right\|$$

$$= \left\|\frac{5}{2}(1, -1, 2)\right\| = \frac{5\sqrt{6}}{2}.$$

$\text{dist}(\mathbf{v}, \mathbb{W})$ is the length $\|\mathbf{v} - P\mathbf{v}\|$ of the vector $\mathbf{v} - P\mathbf{v}$ and it is the distance from point $(-5, 4, 12)$ to the plane $\mathbb{W}$ (see Fig. 1.3).   ∎

Example 9 Let $f(x)$ and $\mathbb{W}$ be the function and subspace of $\widetilde{L}_2[0, \pi]$ considered in Example 7. Compute the distance, $\text{dist}(f, \mathbb{W})$, of f to $\mathbb{W}$.

Solution It follows from the definition (1.4.9) of the distance $\text{dist}(f, \mathbb{W})$ of f to $\mathbb{W}$ that

$$\left(\text{dist}(f, \mathbb{W})\right)^2 = \|f\|^2 - \sum_{j=1}^{3} \frac{|\langle f, g_j \rangle|^2}{\|g_j\|^2}$$

$$= \int_0^{\pi} x^4 dx - \left(\frac{1}{\sqrt{\pi}} \frac{\pi^3}{3}\right)^2 - \left(\frac{\sqrt{2}}{\sqrt{\pi}}(-2\pi)\right)^2 - \left(\frac{\sqrt{2}}{\sqrt{\pi}} \frac{\pi}{2}\right)^2$$

$$= \frac{4}{45}\pi^5 - \frac{17}{2}\pi,$$

so that $\text{dist}(f, \mathbb{W}) = \sqrt{\frac{4}{45}\pi^5 - \frac{17}{2}\pi}$.   ∎

The importance of Bessel's inequality is that the inequality remains valid for all n-dimensional subspaces of $\mathbb{V}$. In particular, if $\mathbb{V}$ is an infinite-dimensional inner-product space, then the inequality (1.4.10) holds even by allowing n to go to infinity. Therefore, for any infinite family $\{\mathbf{w}_1, \mathbf{w}_2, \ldots\}$ of non-zero orthogonal vectors in an infinite-dimensional inner-product space $\mathbb{V}$, we have the following formulation of Bessel's inequality:

$$\sum_{j=1}^{\infty} \frac{|\langle \mathbf{v}, \mathbf{w}_j \rangle|^2}{\|\mathbf{w}_j\|^2} \leq \|\mathbf{v}\|^2. \tag{1.4.11}$$

In the following, we replace $\mathbf{w}_j$ by $\mathbf{w}_j/\|\mathbf{w}_j\|$ in (1.4.11), and show that the inequality in (1.4.11) becomes equality if and only if the normalized family $\{\mathbf{w}_1, \mathbf{w}_2, \ldots\}$ is an orthonormal basis for $\mathbb{V}$.

Theorem 4 $\boxed{\textbf{Orthonormal basis} \Leftrightarrow \textbf{Parseval's identity}}$ *An orthonormal set $W = \{\mathbf{w}_1, \mathbf{w}_2, \ldots\}$ in an inner-product space $\mathbb{V}$ is complete in $\mathbb{V}$, or equivalently an orthonormal basis for $\mathbb{V}$, if and only if it satisfies Parseval's identity:*

$$\sum_{j=1}^{\infty} |\langle \mathbf{v}, \mathbf{w}_j \rangle|^2 = \|\mathbf{v}\|^2, \quad \textit{for all } \mathbf{v} \in \mathbb{V}. \tag{1.4.12}$$

Proof To prove (1.4.12), we observe that since W is an orthonormal basis of $\mathbb{V}$, any $\mathbf{v} \in \mathbb{V}$ can be written as

$$\mathbf{v} = \sum_{j=1}^{\infty} c_j \mathbf{w}_j,$$

with $c_j = \langle \mathbf{v}, \mathbf{w}_j \rangle$, $j = 1, 2, \ldots$ (see Exercise 11). In other words,

$$
\begin{aligned}
0 &= \left\| \mathbf{v} - \sum_{j=1}^{\infty} \langle \mathbf{v}, \mathbf{w}_j \rangle \mathbf{w}_j \right\|^2 \\
&= \lim_{n \to \infty} \left\| \mathbf{v} - \sum_{j=1}^{n} \langle \mathbf{v}, \mathbf{w}_j \rangle \mathbf{w}_j \right\|^2 \\
&= \lim_{n \to \infty} \|\mathbf{v}\|^2 - \sum_{j=1}^{n} |\langle \mathbf{v}, \mathbf{w}_j \rangle|^2 \\
&= \|\mathbf{v}\|^2 - \sum_{j=1}^{\infty} |\langle \mathbf{v}, \mathbf{w}_j \rangle|^2;
\end{aligned}
$$

yielding Parseval's identity (1.4.12).

To prove the converse, suppose that $\mathbf{v} \in \mathbb{V}$. Then for any integer $n > 0$, it follows from (1.4.8) that

$$
\begin{aligned}
\left\| \mathbf{v} - \sum_{j=1}^{n} \langle \mathbf{v}, \mathbf{w}_j \rangle \mathbf{w}_j \right\|^2 \\
= \|\mathbf{v}\|^2 - \sum_{j=1}^{n} |\langle \mathbf{v}, \mathbf{w}_j \rangle|^2 \\
\to \|\mathbf{v}\|^2 - \sum_{j=1}^{\infty} |\langle \mathbf{v}, \mathbf{w}_j \rangle|^2 \text{ as } n \to \infty \\
= 0,
\end{aligned}
$$

by (1.4.12). That is, $\mathbf{v} = \sum_{j=1}^{\infty} \langle \mathbf{v}, \mathbf{w}_j \rangle \mathbf{w}_j$. Since this holds for all $\mathbf{v} \in \mathbb{V}$, W is complete in $\mathbb{V}$ (see Definition 3). This, together with Remark 3, implies that W is an orthonormal basis of $\mathbb{V}$. ∎

To conclude this section, we derive a procedure, called the Gram-Schmidt (orthogonalization) process, for converting a linearly independent family of vectors in an

inner-product space to an orthonormal family, while preserving the same linear span. We will also give three examples for demonstrating this orthogonalization procedure.

Gram-Schmidt process   *Let $U = \{\mathbf{u}_1, \mathbf{u}_2, \ldots \mathbf{u}_n\}$ be a linearly independent set of vectors in an inner-product space $\mathbb{V}$, with inner product $\langle \cdot, \cdot \rangle$. The formulation of the Gram-Schmidt process for obtaining an orthonormal set $V = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ with $\operatorname{span} V = \operatorname{span} U$ is as follows.*

$$\mathbf{v}_1 = \mathbf{u}_1,$$

$$\mathbf{v}_2 = \mathbf{u}_2 - \frac{\langle \mathbf{u}_2, \mathbf{v}_1 \rangle}{\|\mathbf{v}_1\|^2} \mathbf{v}_1,$$

$$\mathbf{v}_3 = \mathbf{u}_3 - \frac{\langle \mathbf{u}_3, \mathbf{v}_1 \rangle}{\|\mathbf{v}_1\|^2} \mathbf{v}_1 - \frac{\langle \mathbf{u}_3, \mathbf{v}_2 \rangle}{\|\mathbf{v}_2\|^2} \mathbf{v}_2,$$

$$\vdots$$

$$\mathbf{v}_n = \mathbf{u}_n - \frac{\langle \mathbf{u}_n, \mathbf{v}_1 \rangle}{\|\mathbf{v}_1\|^2} \mathbf{v}_1 - \frac{\langle \mathbf{u}_n, \mathbf{v}_2 \rangle}{\|\mathbf{v}_2\|^2} \mathbf{v}_2 - \cdots - \frac{\langle \mathbf{u}_n, \mathbf{v}_{n-1} \rangle}{\|\mathbf{v}_{n-1}\|^2} \mathbf{v}_{n-1}.$$

Then the vectors

$$\mathbf{w}_1 = \frac{\mathbf{v}_1}{\|\mathbf{v}_1\|}, \ \mathbf{w}_2 = \frac{\mathbf{v}_2}{\|\mathbf{v}_2\|}, \ \ldots, \ \mathbf{w}_n = \frac{\mathbf{v}_n}{\|\mathbf{v}_n\|}$$

constitute an orthonormal basis for $\operatorname{span} U$. ∎

Observe that for each integer $j > 1$, the new vector $\mathbf{v}_j$ is the difference of $\mathbf{u}_j$ and the orthogonal projection of $\mathbf{u}_j$ onto the subspace $\operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_{j-1}\}$. Thus by statement (c) of Theorem 1, $\mathbf{v}_j$ is orthogonal to $\mathbf{v}_1, \ldots, \mathbf{v}_{j-1}$.

In particular, if $U = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$ is a basis of $\mathbb{V}$, the set $V = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$, obtained by following the Gram-Schmidt process, is an orthogonal basis for $\mathbb{V}$.

Remark 4 In the above formulation of the Gram-Schmidt process, observe that the vector $\mathbf{v}_{j+1}$ does not change even if the previous vectors $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_j\}$ are replaced by $\{c_1\mathbf{v}_1, c_2\mathbf{v}_2, \ldots, c_j\mathbf{v}_j\}$ for arbitrarily chosen non-zero constants $\{c_1, c_2, \ldots, c_j\}$. This should reduce computational complexity if multiplication of $\mathbf{v}_j$ by some appropriate positive integer changes the fractional numerical expression of $\mathbf{v}_j$ to integer representation for computing the norm of $c_j\mathbf{v}_j$ for the formulation of $\mathbf{v}_k$ for $k > j$. For example, if $\mathbf{v}_2 = \frac{1}{5}(-6, 3, 0, 5)$, then we may replace it by $5\mathbf{v}_2 = (-6, 3, 0, 5)$ to simplify the formulation of $\mathbf{v}_j$ for $j > 2$.

The Gram-Schmidt process can be extended to orthogonalization of an infinite set of linearly independent vectors. In particular, if $U = \{\mathbf{u}_1, \mathbf{u}_2, \ldots\}$ is a basis of an infinite-dimensional inner-product space $\mathbb{V}$, then the Gram-Schmidt process can be followed to obtain an orthonormal basis $V = \{\mathbf{v}_1, \mathbf{v}_2, \ldots\}$ of $\mathbb{V}$. This will be an important tool for computing orthonormal families in the later chapters of this book. ∎

Example 10 Find an orthonormal basis of $\operatorname{span}\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$, where

$$\mathbf{u}_1 = (1, 2, 0, 0),$$
$$\mathbf{u}_2 = (-1, 1, 0, 1),$$
$$\mathbf{u}_3 = (1, 0, 2, 1).$$

Solution Following the Gram-Schmidt process, we first set $\mathbf{v}_1 = \mathbf{u}_1$ and compute

$$\|\mathbf{v}_1\|^2 = 1^2 + 2^2 + 0^2 + 0^2 = 5.$$

This yields

$$\mathbf{v}_2 = \mathbf{u}_2 - \frac{\langle \mathbf{u}_2, \mathbf{v}_1 \rangle}{\|\mathbf{v}_1\|^2} \mathbf{v}_1 = \mathbf{u}_2 - \frac{1}{5} \langle (-1, 1, 0, 1), (1, 2, 0, 0) \rangle \mathbf{v}_1$$

$$= (-1, 1, 0, 1) - \frac{1}{5}(1)(1, 2, 0, 0) = \frac{1}{5}(-6, 3, 0, 5).$$

Now, following Remark 4, we compute $\|5\mathbf{v}_2\|^2$ instead of $\|\mathbf{v}_2\|^2$, to obtain

$$\|5\mathbf{v}_2\|^2 = \|(-6, 3, 0, 5)\|^2 = 70.$$

Then it follows from the formula for $\mathbf{v}_3$ in the statement of the Gram-Schmidt process that

$$\mathbf{v}_3 = \mathbf{u}_3 - \frac{\langle \mathbf{u}_3, \mathbf{v}_1 \rangle}{\|\mathbf{v}_1\|^2} \mathbf{v}_1 - \frac{\langle \mathbf{u}_3, 5\mathbf{v}_2 \rangle}{\|5\mathbf{v}_2\|^2} 5\mathbf{v}_2$$

$$= \mathbf{u}_3 - \frac{1}{5} \langle (1, 0, 2, 1), (1, 2, 0, 0) \rangle \mathbf{v}_1 - \frac{1}{70} \langle (1, 0, 2, 1), (-6, 3, 0, 5) \rangle 5\mathbf{v}_2$$

$$= (1, 0, 2, 1) - \frac{1}{5}(1)(1, 2, 0, 0) - \frac{1}{70}(-1)(-6, 3, 0, 5)$$

$$= \frac{1}{70}(50, -25, 140, 75) = \frac{1}{14}(10, -5, 28, 15).$$

This yields the orthogonal basis $\mathbf{v}_1, 5\mathbf{v}_2, 14\mathbf{v}_3$ of span$\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$. Finally, the desired orthonormal basis $\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3$ is obtained by normalization of each of $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ to have unit length, namely:

$$\mathbf{w}_1 = \frac{\mathbf{v}_1}{\|\mathbf{v}_1\|} = \frac{1}{\sqrt{5}}(1, 2, 0, 0),$$

$$\mathbf{w}_2 = \frac{\mathbf{v}_2}{\|\mathbf{v}_2\|} = \frac{5\mathbf{v}_2}{\|5\mathbf{v}_2\|} = \frac{1}{\sqrt{70}}(-6, 3, 0, 5),$$

$$\mathbf{w}_3 = \frac{\mathbf{v}_3}{\|\mathbf{v}_3\|} = \frac{14\mathbf{v}_3}{\|14\mathbf{v}_3\|} = (10, -5, 28, 15)/\|(10, -5, 28, 15)\|$$

$$= \frac{1}{9\sqrt{14}}(10, -5, 28, 15).$$

■

Example 11 Find an orthonormal basis of span$\{x_1, x_2, x_3\}$, where

$$x_1 = (3, 4, 0, 0),$$
$$x_2 = (0, 1, 1, 0),$$
$$x_3 = (0, 0, -3, 4).$$

Solution Since $\langle x_1, x_3 \rangle = 0$, we choose $u_1 = x_1$, $u_2 = x_3$ and $u_3 = x_2$ to save one computational step. In doing so, we already have $v_1 = u_1 = x_1$, $v_2 = u_2 = x_3$ without any work, so that

$$\|v_1\|^2 = 25, \quad \|v_2\|^2 = 25.$$

Then following the Gram-Schmidt procedure, we obtain

$$
\begin{aligned}
v_3 &= u_3 - \frac{\langle u_3, v_1 \rangle}{\|v_1\|^2} v_1 - \frac{\langle u_3, v_2 \rangle}{\|v_2\|^2} v_2 \\
&= x_2 - \frac{1}{25}\langle(0, 1, 1, 0), (3, 4, 0, 0)\rangle v_1 - \frac{1}{25}\langle(0, 1, 1, 0), (0, 0, -3, 4)\rangle v_2 \\
&= (0, 1, 1, 0) - \frac{4}{25}(3, 4, 0, 0) + \frac{3}{25}(0, 0, -3, 4) \\
&= \frac{1}{25}(-12, 9, 16, 12).
\end{aligned}
$$

What remains is only to normalize the vector v_3, yielding the orthonormal basis $\{w_1, w_2, w_3\}$ of span$\{x_1, x_2, x_3\}$, where

$$
\begin{aligned}
w_1 &= \frac{v_1}{\|v_1\|} = \frac{1}{5}(3, 4, 0, 0), \\
w_2 &= \frac{v_2}{\|v_2\|} = \frac{1}{5}(0, 0, -3, 4), \\
w_3 &= \frac{v_3}{\|v_3\|} = \frac{25v_3}{\|25v_3\|} = \frac{(-12, 9, 16, 12)}{\sqrt{144 + 81 + 256 + 144}} \\
&= \frac{1}{25}(-12, 9, 16, 12).
\end{aligned}
$$

Here, as pointed out in Remark 4, we have ignored the multiple $\dfrac{1}{25}$ in the normalization of v_3, since it can be replaced by $25v_3$. ■

Example 12 Let $\mathbb{W} = \text{span}\{f_1, f_2, f_3\} \subset \tilde{L}_2[-1, 1]$, where $f_1(x) = 1$, $f_2(x) = x$, $f_3(x) = x^2$. Find an orthonormal basis of $\mathbb{W}$ by following the Gram-Schmidt process.

Solution Following the Gram-Schmidt process, we set $g_1(x) = f_1(x) = 1$. Then since

$$\langle f_2, g_1 \rangle = \int_{-1}^{1} x \, dx = 0,$$

the two functions g_1 and g_2, with $g_2(x) = f_2(x) = x$, are orthogonal to each other. To obtain g_3, we first compute

$$\langle f_3, g_1 \rangle = \int_{-1}^{1} x^2 \, dx = \frac{2}{3},$$

$$\langle f_3, g_2 \rangle = \int_{1}^{1} x^3 \, dx = 0,$$

$$\|g_1\|^2 = \int_{-1}^{1} 1 \, dx = 2,$$

$$\|g_2\|^2 = \int_{-1}^{1} x^2 \, dx = \frac{2}{3}.$$

Then by the Gram-Schmidt process, we have

$$g_3(x) = f_3(x) - \frac{\langle f_3, g_1 \rangle}{\|g_1\|^2} g_1(x) - \frac{\langle f_3, g_2 \rangle}{\|g_2\|^2} g_2(x)$$

$$= x^2 - \frac{1}{3} g_1(x) - 0 \times g_2(x) = x^2 - \frac{1}{3}.$$

Finally, the desired orthonormal basis h_1, h_2, h_3 of $\mathbb{W}$ is obtained by normalization of each of $\{g_1, g_2, g_3\}$ to have unit length, namely:

$$h_1(x) = \frac{g_1(x)}{\|g_1\|} = \frac{1}{\sqrt{2}},$$

$$h_2(x) = \frac{g_2(x)}{\|g_2\|} = \frac{\sqrt{3}}{\sqrt{2}} x,$$

$$h_3(x) = \frac{g_3(x)}{\|g_3\|} = \frac{3x^2 - 1}{\sqrt{\int_{-1}^{1}(3x^2 - 1)^2 dx}} = \frac{\sqrt{5}}{2\sqrt{2}}(3x^2 - 1).$$

■

Exercises

Exercise 1   Let $\{\mathbf{x}_j\}$ be a (finite or infinite) set of non-zero and mutually orthogonal vectors in an inner-product space $\mathbb{V}$ over some scalar field $\mathbb{F}$. Show that $\{\mathbf{x}_j\}$ is a linearly independent set as defined in Definition 4.

Exercise 2   Determine which of the following sets of vectors in $\mathbb{R}^2$ or $\mathbb{R}^3$ are linearly dependent or linearly independent:

(a) $\mathbf{x}_1 = (1, 3)$, $\mathbf{x}_2 = (4, 1)$, $\mathbf{x}_3 = (-1, 5)$,

(b) $\mathbf{y}_1 = (0, 1, 2)$, $\mathbf{y}_2 = (1, 2, -2)$,
(c) $\mathbf{z}_1 = (1, 0, 1)$, $\mathbf{z}_2 = (0, 1, 1)$, $\mathbf{z}_3 = (1, 1, 0)$,
(d) $\mathbf{w}_1 = (-1, -2, 3)$, $\mathbf{w}_2 = (1, 2, 3)$, $\mathbf{w}_3 = (2, 5, 7)$, $\mathbf{w}_4 = (0, 1, 0)$.

Exercise 3   As a continuation of Exercise 2, determine the dimensions of the following linear spans:

(a) $\mathrm{span}\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$,
(b) $\mathrm{span}\{\mathbf{y}_1, \mathbf{y}_2\}$,
(c) $\mathrm{span}\{\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3\}$,
(d) $\mathrm{span}\{\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3, \mathbf{w}_4\}$.

Exercise 4   Let $n \geq 2$ be an integer. Show that the set $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$ introduced in (1.4.2) is an orthonormal basis of $\mathbb{C}^n$ over $\mathbb{C}$, and also of $\mathbb{R}^n$ over $\mathbb{R}$.

Exercise 5   Fill in the details in the solution of Example 2 by showing that

$$\|\mathbf{x} - \sum_{|k| \leq N} x_k \boldsymbol{\delta}_j\|_2^2 = \sum_{|k| > N} |x_k|^2$$

and that if $\mathbf{x} = \{x_k\} \in \ell_2$, then

$$\sum_{|k| > N} |x_k|^2 \to 0, \quad \text{as } N \to \infty.$$

Exercise 6   Verify that the infinite set $\left\{ \dfrac{1}{\sqrt{\pi}}, \dfrac{\sqrt{2}}{\sqrt{\pi}} \cos jx : j = 1, 2, \ldots \right\}$ is an orthonormal basis of its linear span, as a subspace of $\widetilde{L}_2[0, \pi]$.
   Hint: Recall the trigonometric identity:

$$\cos \alpha \cos \beta = \frac{1}{2} \big(\cos(\alpha - \beta) + \cos(\alpha + \beta) \big).$$

Exercise 7   Verify that the infinite set $\left\{ \dfrac{\sqrt{2}}{\sqrt{\pi}} \sin jx : j = 1, 2, \ldots \right\}$ is an orthonormal basis of its linear span, as a subspace of $\widetilde{L}_2[0, \pi]$.
   Hint: Recall the trigonometric identity:

$$\sin \alpha \sin \beta = \frac{1}{2} \big(\cos(\alpha - \beta) - \cos(\alpha + \beta) \big).$$

Exercise 8   Identify which pairs of bases of $\mathbb{C}^2$ or $\mathbb{R}^2$ in the following are dual bases. Justify your answers.

(a) $\{(1, 1), (1, -1)\}$, $\{(1, 1), (\frac{1}{2}, \frac{-1}{2})\}$.
(b) $\{(1, 1), (1, -1)\}$, $\{(\frac{1}{2}, \frac{1}{2}), (\frac{1}{2}, \frac{-1}{2})\}$.

(c) $\{(1, 1 + i), (0, i)\}, \{(1, 0), (1 + i, -1)\}$.

(d) $\{(1, 1 + i), (0, 1)\}, \{(1, 0), (-1 - i, 1)\}$.

Exercise 9   Let $S = \{\mathbf{v}_k : k = 1, 2, \ldots\}$ be a set of nonzero vectors $\mathbf{v}_k$ in an inner-product space $\mathbb{V}$. Show that if S is orthogonal, then it is linearly independent.

Exercise 10   Prove (a) and (b) Theorem 1.

Exercise 11   Let $W = \{\mathbf{w}_1, \mathbf{w}_2, \ldots\}$ be an orthonormal basis for $\mathbb{V}$. Suppose $\mathbf{v} \in \mathbb{V}$ is represented with W as $\mathbf{v} = \sum_{j=1}^{\infty} c_j \mathbf{w}_j$. Show that $c_j = \langle \mathbf{v}, \mathbf{w}_j \rangle, j = 1, 2, \ldots$.

Exercise 12   Let $\mathbf{w}_1 = (1, 1, 2), \mathbf{w}_2 = (1, 1, -1)$ and $\mathbb{W} = \mathrm{span}\{\mathbf{w}_1, \mathbf{w}_2\}$. For $\mathbf{v} = (-3, 1, 2)$, find the orthogonal projection $P\mathbf{v}$ of $\mathbf{v}$ onto $\mathbb{W}$ and compute the distance $\mathrm{dist}(\mathbf{v}, \mathbb{W})$ of $\mathbf{v}$ to $\mathbb{W}$.

Exercise 13   Repeat Exercise 12 with $\mathbf{w}_1 = (1, 1, 0, 1), \mathbf{w}_2 = (1, 2, -1, -3)$ and $\mathbf{v} = (1, 4, 2, -3)$.

Exercise 14   Let $g_1(x) = \dfrac{1}{\sqrt{\pi}}, g_2(x) = \dfrac{\sqrt{2}}{\sqrt{\pi}} \cos x, g_3(x) = \dfrac{\sqrt{2}}{\sqrt{\pi}} \cos 2x$ be ortho-normal functions in $\widetilde{L}_2[0, \pi]$. Denote $\mathbb{W} = \mathrm{span}\{g_1(x), g_2(x), g_3(x)\}$. Find the orthogonal projection Pf of $f(x) = 1 + x$ onto $\mathbb{W}$ and compute the distance $\mathrm{dist}(f, \mathbb{W})$ of f to $\mathbb{W}$.

Exercise 15   Repeat Exercise 14 with $f(x) = \sin 3x$.

Exercise 16   Repeat Exercise 14 with $g_1(x) = \dfrac{\sqrt{2}}{\sqrt{\pi}} \sin x, g_2(x) = \dfrac{\sqrt{2}}{\sqrt{\pi}} \sin 2x,$ $g_3(x) = \dfrac{\sqrt{2}}{\sqrt{\pi}} \sin 3x$ and $f(x) = 1 + x$.

Exercise 17   Repeat Exercise 14 with $g_1(x) = \dfrac{\sqrt{2}}{\sqrt{\pi}} \sin x, g_2(x) = \dfrac{\sqrt{2}}{\sqrt{\pi}} \sin 2x,$ $g_3(x) = \dfrac{\sqrt{2}}{\sqrt{\pi}} \sin 3x$ and $f(x) = \cos 4x$.

Exercise 18   Let $\mathbf{u}_1, \mathbf{u}_2 \in \mathbb{R}^3$ be given by

$$\mathbf{u}_1 = (0, 0, -1, 1), \quad \mathbf{u}_2 = (1, 0, 0, 1).$$

Find an orthonormal basis of $\mathrm{span}\{\mathbf{u}_1, \mathbf{u}_2\}$ by following the Gram-Schmidt process.

Exercise 19   Let $\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3 \in \mathbb{R}^4$ be given by

$$\mathbf{w}_1 = (0, 0, -1, 1), \quad \mathbf{w}_2 = (1, 0, 0, 1), \quad \mathbf{w}_3 = (0, 1, 1, 0)$$

and $\mathbb{V} = \mathrm{span}\{\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3\}$. Show that $\dim \mathbb{V} = 3$ by verifying that $\mathbf{w}_1, \mathbf{w}_2,$ and $\mathbf{w}_3$ are linearly independent. Then find an orthonormal basis of $\mathbb{V}$.

Hint: Apply the Gram-Schmidt process to $\mathbf{u}_1 = \mathbf{w}_2$, $\mathbf{u}_2 = \mathbf{w}_3$ and $\mathbf{u}_3 = \mathbf{w}_1$.

Exercise 20 Let $\mathbb{W} = \text{span}\{f_1, f_2, f_3\}$ be a subspace of $\widetilde{L}_2[0, 1]$, where $f_1(x) = 1$, $f_2(x) = x$, $f_3(x) = x^2$. Find an orthonormal basis of $\mathbb{W}$ by following the Gram-Schmidt process.

Exercise 21 Let J be an interval in $\mathbb{R}$ and $w(x) \geq 0$, with $w(x) \not\equiv 0$ on any subinterval of J. Define the inner-product space $\widetilde{L}_2(J, w(x)dx)$ of functions $f \in PC(J)$ with $\int_J |f(x)|^2 w(x)dx < \infty$, and with inner product $\langle f, g \rangle = \int_J f(x)\overline{g(x)}w(x)dx$. For $J = [0, \infty)$ and $w(x) = e^{-x}$, follow the Gram-Schmidt process to compute an orthonormal basis of $\mathbb{W} = \text{span}\{f_1, f_2, f_3\} \in \widetilde{L}_2([0, \infty), e^{-x}dx)$, where $f_1(x) = 1$, $f_2(x) = x$, $f_3(x) = x^2$.

1.5 Metric Spaces and Completion

In this section we introduce the notion of metric and normed spaces, and discuss the concept of "complete metric spaces". In particular, the completion of $\widetilde{L}_p(J)$, by extending the piecewise continuous functions $\widetilde{L}_p(J)$ to include all measurable functions, will be denoted by $L_p(J)$.

Definition 1 ⏐**Metric space**⏐ *A set U is called a metric space if there exists a function $d(\mathbf{x}, \mathbf{y})$, called metric (or "distance measurement"), defined for all $\mathbf{x}, \mathbf{y} \in U$, that has the following properties.*

(a) Positivity: $d(\mathbf{x}, \mathbf{y}) \geq 0$ *for all* $\mathbf{x}, \mathbf{y} \in U$, *and* $d(\mathbf{x}, \mathbf{y}) = 0 \Leftrightarrow \mathbf{x} = \mathbf{y}$.
(b) Symmetry: $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$ *for all* $\mathbf{x}, \mathbf{y} \in U$.
(c) Triangle inequality:

$$d(\mathbf{x}, \mathbf{y}) \leq d(\mathbf{x}, \mathbf{z}) + d(\mathbf{z}, \mathbf{y}) \quad \textit{for all } \mathbf{x}, \mathbf{y}, \mathbf{z} \in U.$$

For example, the set $\mathbb{R}$ of real numbers is a metric space with metric defined by the absolute value, namely: $d(x, y) = |x - y|$. More generally, for any positive integer n, the Euclidean space $\mathbb{R}^n$ is a metric space with metric defined by

$$d(\mathbf{x}, \mathbf{y}) = \left((x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2\right)^{\frac{1}{2}}.$$

It is important to point out that a metric space is only a set, which may not be a vector space. In fact, the concept of scalar field is not required in the definition of metric spaces. As an example, any bounded interval $J \in \mathbb{R}$ is a metric space, with metric defined by the absolute value as the metric space $\mathbb{R}$ as above, though J is not closed under addition. Another example of metric spaces that are not vector spaces is the set $U[a, b]$ of continuous functions f on an interval $[a, b]$ with $f(a) = 1$ with metric

defined by

$$d(f, g) = \max_{x \in [a,b]} |f(x) - g(x)|, \quad f, g \in U[a, b]. \tag{1.5.1}$$

Note that $0 = 0 \times f(a) \neq 1$. Thus $g(x) = 0f(x) = 0 \notin U[a, b]$ since $g(a) \neq 1$.

Theorem 1 $\boxed{\ell_p \text{ is metric space with } \| \cdot \|_p}$ *For each $p, 0 \leq p \leq \infty$, the vector space ℓ_p (with $\| \cdot \|_p$ as defined by* (1.2.1)–(1.2.4) *on p.14) is a metric space with metric $d_p(\mathbf{x}, \mathbf{y})$ defined by*

$$d_p(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_p, \quad \mathbf{x}, \mathbf{y} \in \ell_p. \tag{1.5.2}$$

Indeed, the positivity and symmetry properties (a) and (b) for the metric d_p are consequences of the following properties of ℓ_p:

$$\|\mathbf{x}\|_p \geq 0 \text{ for } \mathbf{x} \in \ell_p; \|\mathbf{x}\|_p = 0 \Leftrightarrow \mathbf{x} = 0; \text{ and } \| - \mathbf{x}\|_p = \|\mathbf{x}\|_p,$$

while the triangle inequality property (c) for d_p follows from the triangle inequality (1.2.5) on p.15, namely:

$$\begin{aligned} d_p(\mathbf{x}, \mathbf{y}) &= \|(\mathbf{x} - \mathbf{z}) + (\mathbf{z} - \mathbf{y})\|_p \\ &\leq \|\mathbf{x} - \mathbf{z}\|_p + \|\mathbf{z} - \mathbf{y}\|_p = d_p(\mathbf{x}, \mathbf{z}) + d_p(\mathbf{z}, \mathbf{y}). \end{aligned}$$

$\blacksquare$

Similarly, for any $0 < p \leq \infty$, it follows from the definition of $\| f \|_p$ in (1.2.15)–(1.2.17) and the triangle inequality (1.2.18) on p.22, that the following result, analogous to Theorem 1 for ℓ_p, is also valid for the function space $\widetilde{L}_p(J)$.

Theorem 2 $\boxed{\widetilde{L}_p \text{ is metric space with } \| \cdot \|_p}$ *For each $p, 0 < p \leq \infty$, $\widetilde{L}_p(J)$ is a metric space with metric defined by*

$$d_p(f, g) = \|f - g\|_p, \quad f, g \in \widetilde{L}_p(J). \tag{1.5.3}$$

Since the proof is analogous to that of Theorem 1, it is safe to leave it as an exercise (see Exercise 3).

We next introduce the notion of normed spaces (also called normed linear spaces).

Definition 2 $\boxed{\text{Normed space}}$ *A vector space $\mathbb{V}$ over some scalar field $\mathbb{F}$ is called a normed space, if there is a function $\| \cdot \|$ defined on $\mathbb{V}$ that has the following properties.*

(a) Positivity: $\|\mathbf{x}\| \geq 0$ *for all* $\mathbf{x} \in \mathbb{V}$, *and* $\|\mathbf{x}\| = 0 \Leftrightarrow \mathbf{x} = \mathbf{0}$.
(b) Scale preservation: $\|a\mathbf{x}\| = |a| \|\mathbf{x}\|$ *for all* $\mathbf{x} \in \mathbb{V}$, $a \in \mathbb{F}$.
(c) Triangle inequality:

$$\|\mathbf{x} + \mathbf{y}\| \le \|\mathbf{x}\| + \|\mathbf{y}\| \quad \text{for all } \mathbf{x}, \mathbf{y} \in \mathbb{V}.$$

For a normed space, the function $\|\cdot\|$ is called the "norm".

Theorem 3 $\boxed{\ell_p, \widetilde{L}_p \text{ are normed spaces with } \|\cdot\|_p \text{ for } 1 \le p \le \infty}$ *For each* p
with $1 \le p \le \infty$, *let* $\|\mathbf{x}\| = \|\mathbf{x}\|_p$ *and* $\|f\| = \|f\|_p$, *where* $\mathbf{x} = \{x_j\} \in \ell_p$
and $f \in \widetilde{L}_p(J)$, *be defined by* (1.2.1)–(1.2.2) *on p.14 and* (1.2.15)–(1.2.16) *on p.20,*
respectively. Then ℓ_p *and* $\widetilde{L}_p(J)$ *are normed spaces. However, for* $0 < p < 1$, ℓ_p
and $\widetilde{L}_p(J)$, *with* $\|\mathbf{x}\|_p$ *and* $\|f\|_p$ *defined by* (1.2.3) *on p.14 and* (1.2.17) *on p.20, are*
not normed spaces.

The proof of this theorem for $1 \le p \le \infty$ follows from the triangle inequalities
(1.2.5) on p.15 and (1.2.18) on p.22, since the scale-preservation condition

$$\begin{cases} \|c\mathbf{x}\|_p = |c| \, \|\mathbf{x}\|_p, \\ \|cf\|_p = |c| \, \|f\|_p \end{cases}$$

is satisfied due to the cancellation of the power p in $|c|^p$ by $1/p$.

On the other hand, for $0 < p < 1$, although the triangle inequality holds for $\|\cdot\|_p$
as defined by (1.2.3) and (1.2.17), the scale-preservation condition is not satisfied
(unless the scale c satisfies $|c| = 1$), without pth root, $(1/p)$th power, to cancel the
pth power in $|c|^p$. In other words, for each p with $0 < p < 1$, ℓ_p and $\widetilde{L}_p(J)$ are **not**
normed spaces.   ∎

The definition of ℓ_p measurement in (1.2.1)–(1.2.3) on p.14 for infinite sequences
is valid for finite sequences, which are nothing but vectors in the Euclidean space
$\mathbb{C}^n$. Precisely, for $\mathbf{x} = (x_1, \ldots, x_n)$ in $\mathbb{C}^n$, we define

(a) if $1 \le p < \infty$,

$$\|\mathbf{x}\|_p = \left(\sum_{j=1}^n |x_j|^p \right)^{1/p}; \tag{1.5.4}$$

(b) if $p = \infty$,

$$\|\mathbf{x}\|_\infty = \max\{|x_j| : j = 1, \ldots, n\}; \tag{1.5.5}$$

(c) if $0 < p < 1$,

$$\|\mathbf{x}\|_p = \sum_{j=1}^n |x_j|^p. \tag{1.5.6}$$

Then by (1.2.5), we have the triangle inequality

$$\|\mathbf{x} + \mathbf{y}\|_p \le \|\mathbf{x}\|_p + \|\mathbf{y}\|_p, \quad \text{for any } \mathbf{x}, \mathbf{y} \in \mathbb{C}^n,$$

which yields the following result.

Theorem 4 $\boxed{\ell_p^n \text{ space}}$ *Let ℓ_p^n be the vector space $\mathbb{C}^n$ over scalar field $\mathbb{C}$ with measurement $\|\cdot\|_p$ defined by (1.5.4)–(1.5.6). Then for each p with $0 < p \leq \infty$, ℓ_p^n is a metric space with metric defined by*

$$d_p(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_p, \quad \mathbf{x}, \mathbf{y} \in \ell_p^n.$$

Moreover, for each p with $1 \leq p \leq \infty$, ℓ_p^n is a normed space with norm given by $\|\cdot\|_p$. Furthermore, the same statement is valid if $\mathbb{C}^n$ is replaced by $\mathbb{R}^n$ and $\mathbb{C}$ by $\mathbb{R}$.

We remark that for the choice of $p = 2$, the metric space ℓ_2^n is the Euclidean space $\mathbb{C}^n$ (or $\mathbb{R}^n$).

Theorem 5 $\boxed{\text{Normed space is metric space}}$ *If $\mathbb{V}$ is a normed vector space with norm $\|\cdot\|$, then $\mathbb{V}$ is a metric space with metric defined by*

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|, \quad \mathbf{x}, \mathbf{y} \in \mathbb{V}.$$

Clearly $d(\mathbf{x}, \mathbf{y})$ defined above satisfies (a) and (b) of metric space. Furthermore, we have

$$\begin{aligned}
d(\mathbf{x}, \mathbf{y}) &= \|(\mathbf{x} - \mathbf{z}) + (\mathbf{z} - \mathbf{y})\| \\
&\leq \|\mathbf{x} - \mathbf{z}\| + \|\mathbf{z} - \mathbf{y}\| = d(\mathbf{x}, \mathbf{z}) + d(\mathbf{z}, \mathbf{y}).
\end{aligned} \tag{1.5.7}$$

That is, the triangle inequality holds for $d(\mathbf{x}, \mathbf{y})$. Therefore, $\mathbb{V}$ is a metric space with metric defined as the distance between two vectors.

However, a metric space which is also a vector space is not necessarily a normed space. For example, the spaces ℓ_p and $\widetilde{L}_p(J)$, for $0 < p < 1$, are not normed spaces, although they are metric spaces and vector spaces.

Theorem 6 $\boxed{\text{Inner-product space is normed space}}$ *If $\mathbb{V}$ is an inner-product space with inner product $\langle \cdot, \cdot \rangle$, then $\mathbb{V}$ is a normed space, with norm $\|\cdot\| = \sqrt{\langle \cdot, \cdot \rangle}$.*

Indeed, the property (a) follows from (c) in the definition of an inner product in Definition 1 on p. 8; while the property (b) follows from (b) in this definition by setting $b = 0$ and $\mathbf{z} = a\mathbf{x}$, namely:

$$\|a\mathbf{x}\|^2 = \langle a\mathbf{x}, a\mathbf{x} \rangle = a\bar{a}\langle \mathbf{x}, \mathbf{x} \rangle = |a|^2 \|\mathbf{x}\|^2$$

so that $\|a\mathbf{x}\| = |a| \|x\|$. To complete the proof of Theorem 6, we simply apply Cauchy-Schwarz's inequality (1.3.8) on p.27 in Sect. 1.3 to derive the triangle inequality (c) in Definition 2, as follows:

$$\begin{aligned}
\|\mathbf{x} + \mathbf{y}\|^2 &= |\langle \mathbf{x} + \mathbf{y}, \mathbf{x} + \mathbf{y} \rangle| \\
&= |\langle \mathbf{x}, \mathbf{x} + \mathbf{y} \rangle + \langle \mathbf{y}, \mathbf{x} + \mathbf{y} \rangle| \\
&\leq |\langle \mathbf{x}, \mathbf{x} + \mathbf{y} \rangle| + |\langle \mathbf{y}, \mathbf{x} + \mathbf{y} \rangle|
\end{aligned}$$

$$\leq \|\mathbf{x}\| \, \|\mathbf{x} + \mathbf{y}\| + \|\mathbf{y}\| \, \|\mathbf{x} + \mathbf{y}\|,$$

so that the triangle inequality $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ is obtained by canceling out $\|\mathbf{x} + \mathbf{y}\|$. Observe that this proof is precisely the same as the derivation of Minkowski's inequality by applying Hölder's inequality in Sect. 1.2.   ∎

We next introduce the concept of "completeness" of a metric space (such as a normed space or an inner-product space). For this purpose, we first recall the notion of Cauchy sequences from Calculus, as follows.

Let $\mathbb{V}$ be a metric space with metric $d(\cdot, \cdot)$. A sequence $\{\mathbf{x}_k\}_{k=1}^{\infty} \subset \mathbb{V}$ is called a Cauchy sequence, if

$$d(\mathbf{x}_m, \mathbf{x}_n) \to 0$$

as $m, n \to \infty$, independently. In other words, for any given positive $\epsilon > 0$, there exists a positive integer N, such that

$$d(\mathbf{x}_m, \mathbf{x}_n) < \epsilon, \quad \text{for all } m, n \geq N.$$

Definition 3 | **Complete metric space** | *A metric space $\mathbb{V}$ is said to be complete with respect to some metric $d(\cdot, \cdot)$, if for every Cauchy sequence $\{\mathbf{x}_k\}$ in $\mathbb{V}$, there exists some $\mathbf{x} \in \mathbb{V}$, called the limit of $\{\mathbf{x}_k\}$, such that*

$$d(\mathbf{x}_k, \mathbf{x}) \to 0, \quad \text{as } k \to \infty.$$

A metric space with metric $d(\cdot, \cdot)$ is said to be incomplete, if it is not a complete metric space with this metric.

The set $\mathbb{R}$ of real numbers, considered as a metric space with metric induced by the absolute value, is complete. Furthermore, any closed and bounded interval $[a, b] \in \mathbb{R}$ is also a complete metric space, but the open and bounded interval $(a, b) \in \mathbb{R}$ is incomplete, since a Cauchy sequence that converges to one of the end-points a or b does not have a limit in (a, b). As to functions, the metric space $C[a, b]$ of continuous functions on $[a, b]$ with metric $d(f, g)$ defined by (1.5.1) is complete. This is a well-known result in Advanced Calculus, which states that the limit of a uniformly convergent sequence of continuous functions on a closed and bounded interval is a continuous function on the interval (see the hint in Exercise 6 for its proof). On the other hand, the fact that the subset $\mathbb{Q} \subset \mathbb{R}$ of rational numbers is incomplete, as already alluded to in Sect. 1.1, is shown in the following.

Example 1 Give an example to show that the set $\mathbb{Q}$ of rational numbers, with metric $d(x, y) = |x - y|, x, y \in \mathbb{R}$, is incomplete.

Solution Let $\{x_n\}$ be the sequence of fractions defined by

$$x_1 = 2, \; x_n = \frac{1}{2}x_{n-1} + \frac{1}{x_{n-1}}, \quad n = 2, 3, \ldots.$$

Being fractions, each x_n, $n = 1, 2, \ldots$, is a positive rational number. In addition, in view of the inequality $a + b \geq 2\sqrt{ab}$ for non-negative real numbers, it follows that

$$x_n = \frac{1}{2}x_{n-1} + \frac{1}{x_{n-1}} \geq 2\sqrt{\left(\frac{1}{2}x_{n-1}\right)\left(\frac{1}{x_{n-1}}\right)} = \sqrt{2},$$

so that for all $n = 2, 3, \ldots$, we have

$$x_n - x_{n-1} = \frac{1}{x_{n-1}} - \frac{1}{2}x_{n-1} = \frac{2 - x_{n-1}^2}{2x_{n-1}} \leq 0, \text{ or } x_n \geq x_{n-1}.$$

Thus, as a non-increasing sequence which is bounded from below by $\sqrt{2}$, $\{x_n\}$ must be convergent, and hence, is a Cauchy sequence, with limit denoted by x. According to the definition of the sequence x_n, we see that $x \geq \sqrt{2}$ and satisfies the equation

$$x = \frac{1}{2}x + \frac{1}{x},$$

or $x^2 = 2$, with positive solution $x = \sqrt{2}$, which is not in $\mathbb{Q}$. This example shows that $\mathbb{Q}$ is incomplete.   ∎

Example 2 $\widetilde{L}_p(J)$ is incomplete for each p, $0 < p \leq \infty$.

Solution For a sequence of functions $f_n \in \widetilde{L}_p(J)$, each function of this sequence, being piecewise continuous, is assumed to have at most a finite number of jump discontinuities. Let us assume that f_n has exactly n jump discontinuities in the open interval (a, b). Then if the sequence f_n has some limit f, it should be at least intuitively clear that the limit function f has infinitely many jumps of discontinuities, and is therefore not in $\widetilde{L}_p(J)$. To be precise, consider the example $J = [0, 1]$ and the sequence of functions

$$f_n(x) = \begin{cases} 1/k^s, & \text{for } \frac{1}{k+1} < x \leq \frac{1}{k}, k = 1, \ldots, n, \\ 0, & \text{for } 0 \leq x \leq \frac{1}{n+1}, \end{cases}$$

for some s with $sp > -1$ when $0 < p < \infty$, and $s > 0$ for $p = \infty$ (for example one may just let $s = 1$), where $n = 1, 2, \ldots$. Hence, we have

(a) for $0 < p < \infty$,

$$\int_0^1 |f_n|^p = \sum_{k=1}^n \frac{1}{k^{ps}k(k+1)} < \infty;$$

(b) for $p = \infty$,

$$\operatorname*{ess\,sup}_{x \in [0,1]} |f_n(x)| = 1.$$

Furthermore, for $0 < p < \infty$ and s with $ps > -1$, we have, for $m > n$,

$$d_p(f_m, f_n) = \| f_m - f_n \|_p = \left(\sum_{k=n+1}^{m} \frac{1}{k^{ps} k(k+1)} \right)^{1/p} \to 0, \qquad (1.5.8)$$

as $m, n \to \infty$ independently.

For $p = \infty$, we also have, for any $s > 0$,

$$d_\infty(f_m, f_n) = \operatorname{ess} \sup_{n < k \leq m} \frac{1}{k^s} = \frac{1}{(n+1)^s} \to 0,$$

as $m, n \to \infty$. Hence, $\{f_n\}$ is a Cauchy sequence in $\widetilde{L}_p(J), 0 < p \leq \infty$.

It should be clear that according to the definition of $f_n(x)$, the "limit function" $f_*(x)$, as $n \to \infty$, is defined by

$$f_*(x) = \begin{cases} 1/k^s, & \text{for } \frac{1}{k+1} < x \leq \frac{1}{k}, \ k = 1, 2, \dots, \\ 0, & \text{for } x = 0. \end{cases}$$

However, observe that the limit function f_* with infinitely many jump discontinuities is not in $\widetilde{L}_p(J)$, for each p, $0 < p \leq \infty$. Hence, the spaces $\widetilde{L}_p(J)$ are not complete. ∎

Remark 1 It is perhaps intuitively clear that to complete the space $\widetilde{L}_p(J)$, the "limit functions" of all Cauchy sequences must be added to $\widetilde{L}_p(J)$. Unfortunately, although this is partially sufficient, there is more to this simple idea, since the integral defined for piecewise continuous functions must also be extended to "measurable" functions. More precisely, the notion of integration from Calculus defined by using Riemann sums must be extended to "Lebesgue integration". To understand the notion of the Lebesgue integral, we recall from Remark 2 in Sect. 1.1 that two functions f and g are said to be equal almost everywhere (in short, $f = g$ a.e.), if $f(x) = g(x)$ for all x outside a set of measure zero. For $f = g$ a.e., with $g \in \widetilde{L}_1(J)$, then we say that f is Lebesgue integrable, with integral equal to the integral of g.

To give an illustration for this statement, consider the function f defined on the interval $J = [0, 1]$ by

$$f(x) = \begin{cases} 1, & \text{for rational } x \in J, \\ 0, & \text{for irrational } x \in J. \end{cases}$$

Then since $f(x) = 0$ a.e. and the constant function 0 has integral equal to 0 on J, the function f defined above is Lebesgue integrable on J, with integral equal to 0, but it is not Riemann integrable. ∎

Remark 2 For $p = \infty$, the notion of "ess sup" introduced in (1.2.16) on p.20 for piecewise continuous functions must also be extended to measurable functions, by

replacing finite sets of jump discontinuities by measurable sets with measure zero (see Remarks 1 and 2 in Sect. 1.1). ∎

Definition 4 | **Complete normed space L_p** | *After extending the Riemann integral to the Lebesgue integral, the completion of $\widetilde{L}_p(J)$ is denoted by $L_p(J)$, for each p, $0 < p \leq \infty$. This is accomplished by taking the closure (or adding the limits of all Cauchy sequences) in $\widetilde{L}_p(J)$, with respect to the metric d_p in (1.5.3), and the notation*

$$L_p(J) = \text{closure}_{d_p} \widetilde{L}_p(J) \tag{1.5.9}$$

will be used.

A complete normed space, such as $L_p(J)$ and ℓ_p with $1 \leq p \leq \infty$, is called a Banach space; and a Banach space with its norm defined by an inner product as in (1.3.7) on p.26 is called a Hilbert space. Therefore, both the function space $L_2(J)$ and the sequence space ℓ_2 are Hilbert spaces.

<div align="center">

Exercises

</div>

Exercise 1   For each of the following, decide if $U = \mathbb{R}^2$ is a metric space with $d(\cdot, \cdot)$ given by:

(a) $d\big((x_1, y_1), (x_2, y_2)\big) = \sqrt{2(x_1 - x_2)^2 + (y_1 - y_2)^2}$,
(b) $d\big((x_1, y_1), (x_2, y_2)\big) = |x_1 - x_2| + 3|y_1 - y_2|$,
(c) $d\big((x_1, y_1), (x_2, y_2)\big) = |x_1 - x_2|$,
(d) $d\big((x_1, y_1), (x_2, y_2)\big) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$.

Exercise 2   Let $w_1, w_2, \ldots, w_n$ be positive numbers. Show that

$$d(\mathbf{x}, \mathbf{y}) = \Big(w_1(x_1 - y_1)^2 + w_2(x_2 - y_2)^2 + \cdots + w_n(x_n - y_n)^2 \Big)^{\frac{1}{2}}$$

is a metric for $\mathbb{R}^n$. Suppose that one of the positive numbers $w_1, w_2, \ldots, w_n$ is replaced by 0, is $d(\mathbf{x}, \mathbf{y})$ still a metric for $\mathbb{R}^n$? Justify your answer by a rigorous argument or a counter-example.

Exercise 3   Let $0 < p \leq \infty$. Follow the proof of Theorem 1 to give a proof of Theorem 2.

Exercise 4   Show that (i) $U = (0, 1)$ is a metric space with metric $d(x, y) = |x - y|$, $x, y \in U$; (ii) this metric space U is incomplete.

Exercise 5   Let x_n, $n = 1, 2, \ldots$, be a sequence of fractions (or rational numbers) defined by

$$x_1 = 3, \ x_n = \frac{1}{2}x_{n-1} + \frac{3}{2x_{n-1}}, \quad n = 2, 3, \ldots .$$

Show that $\{x_n\}$ is convergent and compute its limit x. Is the limit x a rational number?

Exercise 6   Show that the metric space $C[a, b]$ of continuous functions on the closed and bounded interval $[a, b]$, with metric $d(\cdot, \cdot)$ defined by (1.5.1), is complete.

Hint: First show that a Cauchy sequence $f_n \in C[a, b]$, with metric $d(\cdot, \cdot)$, converges for each fixed $x \in [a, b]$ to some limit $f(x)$, so that f is a function defined on $[a, b]$. Next, show that $d(f_n, f) \to 0$ as $n \to \infty$ implies that the convergence of $f_n(x)$ to $f(x)$ is uniform for $x \in [a, b]$. Finally prove that the limit function f of the uniformly convergent sequence f_n of continuous functions on $[a, b]$ is also a continuous function on $[a, b]$.

Exercise 7   Decide if ℓ_0 is a complete metric space. Give a rigorous argument or a counter-example to justify your answer.

Exercise 8   Show that for each real number p with $1 < p < \infty$, ℓ_p is a complete metric space.

Exercise 9   Decide if ℓ_∞ is a complete space. Give a rigorous argument or a counter-example to justify your answer.

Exercise 10   Let $\{\mathbf{v}_k : k = 0, \pm 1, \pm 2, \ldots\}$ be an orthonormal family of sequences in ℓ_2. Show that though the set $\{\|\mathbf{v}_k\|_2 : k = 0, \pm 1, \pm 2, \ldots\}$ is bounded, any subsequence of $\{\mathbf{v}_k\}_{k=\ldots,-1,0,1,\ldots}$ is divergent.

Hint: Calculate $\|\mathbf{v}_j - \mathbf{v}_k\|_2$, and apply the result from your computation to show that any subsequence of $\{\mathbf{v}_k\}_{k=\ldots,-1,0,1,\ldots}$ is not a Cauchy sequence.

Exercise 11   Show that the infinite series

$$\sum_{k=1}^{\infty} \frac{1}{k^\alpha}$$

is convergent if and only if $\alpha > 1$.

Hint: Show that

$$\int_1^{N+1} \frac{dx}{x^\alpha} < \sum_{k=1}^{N} \frac{1}{k^\alpha} < 1 + \int_1^{N} \frac{dx}{x^\alpha},$$

and apply these inequalities to study the convergence or divergence of the infinite series.

Exercise 12   Apply the result in Exercise 11 to verify (1.5.8).

Chapter 2
Linear Analysis

$$\mathbf{A} = \mathbf{U} \begin{bmatrix} \lambda_1 & & & & & \\ & \ddots & & & & \\ & & \lambda_r & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{bmatrix} \mathbf{U}^*$$

In the previous chapter, the theory and methods of elementary linear algebra were extended to linear spaces, in that finite-dimensional vector spaces were extended to include infinite-dimensional sequence and function spaces, that the inner product was applied to derive such mathematical tools as orthogonal projection and the Gram-Schmidt orthogonalization process, and that the concept of completeness was studied to introduced Hilbert and Banach spaces. In this chapter, the basic theory of linear analysis is studied in some depth with the goal of preparing for the later discussion of several application areas, including: solution and least-squares approximation for linear systems and reduction of data dimension (in Chap. 3) and image/video compression (in Chap. 5), as well as for the study of anisotropic diffusion and compressed sensing in a forthcoming publication of this book series.

In the first section, Sect. 2.1, we give a brief review of the basic theory and methods of elementary linear algebra, including matrix-matrix multiplication, block matrix operations, row echelon form, and matrix rank. An application to the study of the rank of a matrix is also discussed at the end of this first section, with the objective of preparing for the study of spectral methods and their applications in the next chapter.

In the next Sect. 2.2, matrix multiplication to column vectors is extended to the notion of linear transformations, particularly linear functionals and linear operators, defined on vector spaces. A representation formula for any bounded linear functional defined on an inner-product space is derived and applied to prove the existence and uniqueness of the adjoint of an arbitrary bounded linear operator on a complete inner-product space. In addition, when a square matrix A of dimension n is considered as a linear operator on the Euclidean space $\mathbb{C}^n$, it is shown that its adjoint is given by complex conjugation of the transpose of A, which is different from the matrix adjoint of A, studied in elementary matrix theory or linear algebra for the inversion of the matrix A, if it happens to be nonsingular.

In Sect. 2.3, the subject of eigenvalues and eigenvectors of square matrices, studied in elementary linear algebra, is reviewed and extended to linear operators on arbitrary vector spaces. To prepare for the solution of the (isotropic) heat diffusion PDE (partial differential equation) in Sect. 6.5 of Chap. 6, an example of the eigenvalue/eigenvector problem for certain differential operators is presented. This example is also used to demonstrate the notion of self-adjoint and positive semi-definite linear operators on inner-product spaces. It is shown that eigenvalues of a self-adjoint linear operator T (meaning that T is its own adjoint) are real numbers, and that these real eigenvalues are non-negative, if the self-adjoint operator T is positive semi-definite, meaning that the inner-product of $T\mathbf{v}$ with $\mathbf{v}$, for an arbitrary $\mathbf{v}$ in the inner-product space, is non-negative. This section ends with a discussion of a certain minimization problem in terms of the Rayleigh quotient for the computation of eigenvalues, that will be useful for numerical solution of the (anisotropic) heat diffusion PDE, and serves as a motivation for the study of the spectral decomposition problem in the next section.

The notion of eigenvalues of square matrices is extended to that of spectra for linear operators in Sect. 2.4. A linear operator is said to be normal, if it commutes with its adjoint. The main result established in this section is the Spectral (Decomposition) Theorem for normal matrices, with application to the derivation of the property of preservation of distances, norms, and angles, under normal matrix transformation. Examples of normal matrices include self-adjoint (Hermitian) matrices and unitary matrices (which are also called orthogonal matrices, for those with real entries).

2.1 Matrix Analysis

Let $A \in \mathbb{C}^{m,n}$ and $B \in \mathbb{C}^{n,p}$. Then their product, $AB \in \mathbb{C}^{m,p}$, is defined by

$$
AB = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ & \dots & \\ a_{m1} & \dots & a_{mn} \end{bmatrix} \begin{bmatrix} b_{11} & \dots & b_{1p} \\ & \dots & \\ b_{n1} & \dots & b_{np} \end{bmatrix}
$$

$$
= \begin{bmatrix} c_{11} & \dots & c_{1p} \\ & \dots & \\ c_{m1} & \dots & c_{mp} \end{bmatrix} = C, \tag{2.1.1}
$$

with

$$
c_{jk} = \sum_{\ell=1}^{n} a_{j\ell} b_{\ell k}.
$$

It is important to emphasize that for the validity of the matrix multiplication $AB = C$, the number n of "columns" of A must be the same as the number n of "rows" of B.

Example 1 Consider

$$A = \begin{bmatrix} 1 & -1 & 0 \\ 2 & 3 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 2 & -1 \\ 0 & 1 & 2 \\ 4 & 0 & 1 \end{bmatrix}.$$

Then AB is well defined but BA does not make sense, since $m = 2, n = 3$ and $p = 3$. Compute AB.

Solution By applying the rule of matrix-matrix multiplication defined in (2.1.1), we have

$$AB = \begin{bmatrix} 1 & -1 & 0 \\ 2 & 3 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & -1 \\ 0 & 1 & 2 \\ 4 & 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \end{bmatrix} = C,$$

where, according to (2.1.1),

$$c_{11} = 1(1) + (-1)(0) + 0(4) = 1$$
$$c_{12} = 1(2) + (-1)(1) + 0(0) = 1$$
$$c_{13} = 1(-1) + (-1)(2) + 0(1) = -3$$
$$c_{21} = 2(1) + 3(0) + 1(4) = 6$$
$$c_{22} = 2(2) + 3(1) + 1(0) = 7$$
$$c_{23} = 2(-1) + 3(2) + 1(1) = 5.$$

Hence,

$$AB = C = \begin{bmatrix} 1 & 1 & -3 \\ 6 & 7 & 5 \end{bmatrix}.$$

∎

Remark 1 It is convenient to identify $\mathbb{C}^{n,1}$ with the Euclidean space $\mathbb{C}^n$ by agreeing on the notation $\mathbb{C}^{n,1} = \mathbb{C}^n$. Recall from (1.1.2) on p.8 that

$$\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{C}^n.$$

Hence, by identifying $\mathbb{C}^n$ with the set $\mathbb{C}^{n,1}$ of $n \times 1$ matrices, $\mathbf{x}$ is also a column vector, namely:

$$\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{C}^{n,1}.$$

Hence-forth, we will therefore use both notations $(x_1, \ldots, x_n)$ and $\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$ for the

same $\mathbf{x} \in \mathbb{C}^n = \mathbb{C}^{n,1}$. In particular, for $A \in \mathbb{C}^{m,n}$ and $\mathbf{x} \in \mathbb{C}^n$, we have, by adopting the convention $\mathbf{x} \in \mathbb{C}^{n,1}$,

$$
\begin{aligned}
A\mathbf{x} &= \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \vdots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \\
&= \begin{bmatrix} a_{11}x_1 + \cdots + a_{1n}x_n \\ \vdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n \end{bmatrix}.
\end{aligned} \tag{2.1.2}
$$

∎

On the other hand, in some occasions, it is more convenient to write a matrix A as a **block matrix** or a **partitioned matrix**, such as

$$
A = [A_1 A_2], \, A = \begin{bmatrix} C_1 \\ C_2 \end{bmatrix}, \, A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},
$$

where $A_1, A_2, C_1, C_2, A_{11}, A_{12}, A_{21}, A_{22}$ are submatrices of A. As an example, the matrices A and B in Example 1 may be written as

$$
A = \begin{bmatrix} 1 & \vdots & -1 & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 2 & \vdots & 3 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & \vdots & 2 & -1 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & \vdots & 1 & 2 \\ 4 & \vdots & 0 & 1 \end{bmatrix}. \tag{2.1.3}
$$

In other words, we have partitioned the matrices A and B as block matrices $A = [A_{jk}]_{1 \le j,k \le 2}$, $B = [B_{jk}]_{1 \le j,k \le 2}$, where

$$
A_{11} = [1], \, A_{12} = [-1 \ 0], \, A_{21} = [2], \, A_{22} = [3 \ 1],
$$
$$
B_{11} = [1], \, B_{12} = [2 \ -1], \, B_{21} = \begin{bmatrix} 0 \\ 4 \end{bmatrix}, \, B_{22} = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}.
$$

Remark 2 ⎵ **Block-matrix multiplication** ⎵ The operation of matrix-matrix multiplication (2.1.1) remains valid for multiplication of block matrices. For example, let

$$
A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \, B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix},
$$

where A_{jk} and B_{jk} are the (j, k)-blocks of A and B with appropriate sizes. Then we have

$$AB = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix} = \begin{bmatrix} A_{11}B_{11} + A_{12}B_{21} & A_{11}B_{12} + A_{12}B_{22} \\ A_{21}B_{11} + A_{22}B_{21} & A_{21}B_{12} + A_{22}B_{22} \end{bmatrix}.$$

Also, even non-square matrices can be written as diagonal block matrices. For example, if A and B are diagonal block matrices, denoted by

$$A = \text{diag}\{A_1, A_2, \ldots, A_s\}, \; B = \text{diag}\{B_1, B_2, \ldots, B_s\},$$

with main diagonal blocks $A_j, B_j, 1 \leq j \leq s$ of appropriate sizes and the other block matrices not on the "diagonal" being zero submatrices, then by applying the above multiplication operation for block matrices, we have

$$AB = \text{diag}\{A_1 B_1, A_2 B_2, \ldots, A_s B_s\}.$$

We emphasize again that this multiplication operation applies, even if A, B are not square matrices. ∎

Observe that by writing any $A \in \mathbb{C}^{m,n}$ as a block matrix

$$A = [\mathbf{a}_1 \; \mathbf{a}_2 \; \cdots \; \mathbf{a}_n],$$

of vector column sub-blocks, with $\mathbf{a}_j$ being the jth column of A, then for any $\mathbf{x} \in \mathbb{C}^n$, the matrix-vector multiplication of A and $\mathbf{x}$ can be written as

$$\begin{aligned} A\mathbf{x} &= [\mathbf{a}_1 \; \mathbf{a}_2 \; \cdots \; \mathbf{a}_n] \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \\ &= x_1 \mathbf{a}_1 + x_2 \mathbf{a}_2 + \cdots + x_n \mathbf{a}_n. \end{aligned} \tag{2.1.4}$$

The matrix partitioning of A into blocks of column-vectors for the reformulation (2.1.2) of matrix-vector multiplication will facilitate our discussions below and in later chapters, particularly in Chap. 4 in the discussion of DFT and DCT.

Example 2 Let A and B be the matrices in Example 1, and consider the partition of these two matrices into block matrices as given in (2.1.3). Compute the product AB by multiplication of block matrices.

Solution In terms of the notation for block matrix multiplication introduced in the above discussion, it follows that

$$C_{11} = A_{11}B_{11} + A_{12}B_{21} = [1][1] + \begin{bmatrix} -1 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 4 \end{bmatrix} = [1],$$

$$C_{12} = A_{11}B_{12} + A_{12}B_{22} = [1]\begin{bmatrix} 2 & -1 \end{bmatrix} + \begin{bmatrix} -1 & 0 \end{bmatrix}\begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 2 & -1 \end{bmatrix} + \begin{bmatrix} -1 & -2 \end{bmatrix} = \begin{bmatrix} 1 & -3 \end{bmatrix},$$

$$C_{21} = A_{21}B_{11} + A_{22}B_{21} = [2][1] + \begin{bmatrix} 3 & 1 \end{bmatrix}\begin{bmatrix} 0 \\ 4 \end{bmatrix} = [6],$$

$$C_{22} = A_{21}B_{12} + A_{22}B_{22} = [2]\begin{bmatrix} 2 & -1 \end{bmatrix} + \begin{bmatrix} 3 & 1 \end{bmatrix}\begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 4 & -2 \end{bmatrix} + \begin{bmatrix} 3 & 7 \end{bmatrix} = \begin{bmatrix} 7 & 5 \end{bmatrix}.$$

Thus, we have

$$\begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix} = \begin{bmatrix} 1 & 1 & -3 \\ 6 & 7 & 5 \end{bmatrix},$$

which agrees with the matrix AB obtained by the matrix-matrix multiplication in Example 1. ∎

Next we recall the notions of transpose and transpose-conjugate of matrices. The **transpose** of an $m \times n$ matrix $A = [a_{jk}]_{1 \le j \le m, 1 \le k \le n}$, denoted by A^T, is the $n \times m$ matrix defined by

$$A^T = [a_{kj}]_{1 \le j \le n, 1 \le k \le m} = \begin{bmatrix} a_{11} & a_{21} & \dots & a_{m1} \\ a_{12} & a_{22} & \dots & a_{m2} \\ \vdots & \vdots & \dots & \vdots \\ a_{1n} & a_{2n} & \dots & a_{mn} \end{bmatrix}.$$

The **transpose-conjugate** of A, denoted by A^*, is defined by

$$A^* = \left(\overline{A} \right)^T, \qquad (2.1.5)$$

where $\overline{A} = [\overline{a}_{jk}]$. For real matrices A, it is clear that $A^* = A^T$. A square matrix A is said to be **symmetric** if $A^T = A$, and **Hermitian** if $A^* = A$. We also recall that

$$\left(A^T \right)^T = A, \ \left(AB \right)^T = B^T A^T,$$
$$\left(A^* \right)^* = A, \ \left(AB \right)^* = B^* A^*,$$

which can be easily verified. For block matrices $A = \begin{bmatrix} B & C \\ D & E \end{bmatrix}$, the above operations can be extended to

$$A^T = \begin{bmatrix} B^T & D^T \\ C^T & E^T \end{bmatrix}, A^* = \begin{bmatrix} B^* & D^* \\ C^* & E^* \end{bmatrix}.$$

∎

In an elementary course of Linear Algebra or Matrix Theory, the method of Gaussian elimination is introduced to solve a system of linear equations, to invert nonsingular matrices, study matrix rank, and so forth. The matrices are then reduced to the so-called "row echelon form" or "reduced row echelon form", defined as follows.

Definition 1 | **Row echelon form** | *A matrix is in* **row echelon form** *if it satisfies the following properties.*

(a) *All nonzero rows appear above any rows of all zeros.*
(b) *The first nonzero entry from the left of a nonzero row is a* 1*, and is called the* **leading 1** *of this row.*
(c) *All entries in each column below a leading* 1 *are zeros.*
   A matrix in row echelon form is said to be in **reduced row echelon form** *if it satisfies the following additional condition*:
(d) *Each leading* 1 *is the only nonzero entry in the column that contains this leading* 1.

For example, the following matrices are in row echelon form, where $*$ denotes any number:

$$\begin{bmatrix} 1 & * & * & * \\ 0 & 0 & 1 & * \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \begin{bmatrix} 0 & 1 & * & * & * & * & * \\ 0 & 0 & 1 & * & * & * & * \\ 0 & 0 & 0 & 0 & 1 & * & * \\ 0 & 0 & 0 & 0 & 0 & 1 & * \end{bmatrix},$$

and in particular, the following two matrices

$$\begin{bmatrix} 1 & * & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \begin{bmatrix} 0 & 1 & 0 & * & 0 & 0 & * \\ 0 & 0 & 1 & * & 0 & 0 & * \\ 0 & 0 & 0 & 0 & 1 & 0 & * \\ 0 & 0 & 0 & 0 & 0 & 1 & * \end{bmatrix}$$

are in reduced row echelon form.

A matrix can be transformed into a matrix in row echelon form or reduced row echelon form by performing finitely many steps of the elementary row operations. There are three permissible elementary row operations on a matrix A, to be called types I-III, defined as follows:

 (i) Type I: Interchange two rows of A.
 (ii) Type II: Multiply a row of A by a nonzero constant.
(iii) Type III: Add a constant multiple of a row of A to another row of A.

For example, to study the solution of system of linear equations

$$A\mathbf{x} = \mathbf{b} \tag{2.1.6}$$

(also called a linear system for simplicity), where A is an $m \times n$ (coefficient) matrix, and $\mathbf{b}$ is a given $m \times 1$ column vector, and $\mathbf{x}$ is the unknown column vector in $\mathbb{C}^n$, let

$\begin{bmatrix} C \vdots \mathbf{d} \end{bmatrix}$ be the matrix in row echelon form or reduced row echelon form obtained by applying elementary row operations on the "augmented matrix" $\begin{bmatrix} A \vdots \mathbf{b} \end{bmatrix}$ of the linear system (2.1.6). Then $C\mathbf{x} = \mathbf{d}$ has the same solutions as the original linear system (2.1.6). Thus, by solving $C\mathbf{x} = \mathbf{d}$, which is easy to solve since C is in row echelon form or reduced row echelon form, we get the solution for (2.1.6). This method is called the Gaussian elimination if $\begin{bmatrix} C \vdots \mathbf{d} \end{bmatrix}$ is in row echelon form, and it is called the Guass-Jordan reduction if $\begin{bmatrix} C \vdots \mathbf{d} \end{bmatrix}$ is in reduced row echelon form.

An $n \times n$ matrix R is called an **elementary matrix** of type I, type II or type III (each is called an elementary matrix) if it is obtained by performing one step of type I, type II, or type III elementary row operation on the identity matrix I_n. For example,

$$R_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}, R_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & c & 0 \\ 0 & 0 & 1 \end{bmatrix}, R_2 = \begin{bmatrix} 1 & d & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

where $c \neq 0$, $d \neq 0$, are type I, type II, and type III elementary matrices, respectively. Let $A = [a_{jk}]$ be a 3×3 matrix. Observe that

$$R_1 A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{31} & a_{32} & a_{33} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}, R_2 A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ ca_{21} & ca_{22} & ca_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix},$$

$$R_3 A = \begin{bmatrix} a_{11} + da_{21} & a_{12} + da_{22} & a_{13} + da_{23} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}.$$

Hence, $R_1 A$ is the matrix obtained by interchanging the second and third rows of A, and this is exactly the same operation applied to I_3 to obtain R_1 (namely, R_1 is the resulting matrix after interchanging the second and third rows of I_3). Similarly, $R_2 A$ and $R_3 A$ are the matrices obtained by performing the corresponding type II and type III elementary row operations on A as they are performed on I_3 to yield R_2 and R_3, respectively. More generally, for an elementary matrix R, we have the following fact:

>   *RA* **is the matrix obtained by performing the corresponding elementary row operation on** A .

Thus, if B is the matrix obtained by performing finitely many steps of elementary row operations on A, then

$$B = EA,$$

where $E = R_k \dots R_2 R_1$ with each R_j an elementary matrix. In addition, E is nonsingular, and

$$A = E^{-1} B,$$

where $E^{-1} = R_1^{-1} R_2^{-1} \dots R_k^{-1}$ with each R_j^{-1} being also an elementary matrix. ∎

For an $m \times n$ matrix A, let $\mathbf{a}_j$ denote the jth column of A, namely:

$$A = [\mathbf{a}_1 \ \mathbf{a}_2 \ \cdots \ \mathbf{a}_n].$$

The **column space** of A is defined by $\text{span}\{\mathbf{a}_1, \dots, \mathbf{a}_n\}$; that is, the vector space spanned by the columns of A. The dimension of this vector space is called **column-rank** of the matrix A. Next we apply the above discussion to find a basis for the column space of A.

To find a basis of vectors among $\mathbf{a}_1, \dots, \mathbf{a}_n$ for $\text{span}\{\mathbf{a}_1, \dots, \mathbf{a}_n\}$, we need to take linear dependence of $\mathbf{a}_1, \dots, \mathbf{a}_n$ into consideration, by investigating the linear system

$$c_1 \mathbf{a}_1 + \cdots + c_n \mathbf{a}_n = \mathbf{0},$$

where $c_1, \dots, c_n \in \mathbb{C}$. From (2.1.4), this linear system has matrix formulation: $A\mathbf{c} = \mathbf{0}$, where $\mathbf{c} = [c_1, \dots, c_n]^T$, with coefficient matrix A.

With elementary row operations, we may transform A into its reduced row echelon form B. If $B\mathbf{c} = \mathbf{0}$ has only the trivial solution, then $\mathbf{a}_1, \dots, \mathbf{a}_n$ are linearly independent and therefore constitute a basis for the column space. In this case B must have n leading 1's and that $n \leq m$. Next we may assume $B\mathbf{c} = \mathbf{0}$ has a nontrivial solution. For simplicity of presentation, we assume that the columns of B with a leading 1 precede any columns without a leading 1. More precisely, B can be written as

$$B = \begin{bmatrix} 1 & 0 & \dots & 0 & b_{1\,r+1} & \dots & b_{1n} \\ 0 & 1 & \dots & 0 & b_{2\,r+1} & \dots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & b_{r\,r+1} & \dots & b_{rn} \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & 0 & \dots & 0 \end{bmatrix}, \tag{2.1.7}$$

where $0 \leq r \leq n$.

Let $\mathbf{b}_j, 1 \leq j \leq n$, denote the jth column of B. Clearly the first r columns $\mathbf{b}_1, \dots, \mathbf{b}_r$ of B are linearly independent. In addition, other columns can be written as a linear combination of them:

$$\mathbf{b}_j = b_{1j} \mathbf{b}_1 + b_{2j} \mathbf{b}_2 + \cdots + b_{rj} \mathbf{b}_r, \, r + 1 \leq j \leq n. \tag{2.1.8}$$

Thus, $\mathbf{b}_1, \dots, \mathbf{b}_r$ form a basis for the column space of B.

From the fact that $A = EB$, where E is a nonsingular matrix which is a product of elementary matrices, we see

$$\mathbf{a}_j = E\mathbf{b}_j, 1 \leq j \leq n.$$

Thus, by multiplying both sides of (2.1.8) with E, we have

$$E\mathbf{b}_j = b_{1j}E\mathbf{b}_1 + b_{2j}\ E\mathbf{b}_2 + \cdots + b_{rj}\ E\mathbf{b}_r,\ r+1 \leq j \leq n,$$

or

$$\mathbf{a}_j = b_{1j}\ \mathbf{a}_1 + b_{2j}\ \mathbf{a}_2 + \cdots + b_{rj}\ \mathbf{a}_r,\ r+1 \leq j \leq n.$$

That is,

$$\mathbf{a}_j \in \text{span } \{\mathbf{a}_1, \ldots, \mathbf{a}_r\},\ r+1 \leq j \leq n.$$

On the other hand, from the linear independence of $\mathbf{b}_1, \ldots, \mathbf{b}_r$ and the relation $\mathbf{b}_j = E\mathbf{a}_j$ between $\mathbf{a}_j$ and $\mathbf{b}_j$, it is easy to see that $\mathbf{a}_1, \ldots, \mathbf{a}_r$ are linearly independent (see Exercise 7). Thus, $\{\mathbf{a}_1, \ldots, \mathbf{a}_r\}$ is a basis of the column space of A. To summarize, we have shown that the column-rank of A is r, which is the number of the leading 1's in B, and the columns of A corresponding to the columns B with a leading 1 form a basis for the column space of A. More generally, we have the following theorem.

Theorem 1 $\boxed{\textbf{Basis for column space}}$ *Let A be an $m \times n$ matrix and B be its row echelon form. Then*

$$column\text{-}rank\ A = \#\ of\ leading\ 1's\ in\ B,$$

and the columns of A corresponding to the columns B with a leading 1 constitute a basis of the column space of A.

For a set of vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ in $\mathbb{C}^m$ or $\mathbb{R}^m$, to find a basis for its span, we only need to apply Theorem 1 to the matrix with $\mathbf{v}_j, 1 \leq j \leq n$ (or $\mathbf{v}_j^T$ if $\mathbf{v}_j$ are row vectors) as its column vectors.

Example 3 Determine some columns of the following matrix A which constitute a basis for its column space:

$$A = \begin{bmatrix} 0 & 1 & 2 & 1 & 3 \\ 2 & 4 & 2 & 0 & 8 \\ 3 & 7 & 5 & 1 & 15 \\ 2 & 3 & 0 & 1 & 3 \end{bmatrix}.$$

Solution We first transform A into its row echelon form.

$$A \to \text{ Multiply 2nd row by } \frac{1}{2}$$

$$
\longrightarrow
\begin{bmatrix}
0 & 1 & 2 & 1 & 3 \\
1 & 2 & 1 & 0 & 4 \\
3 & 7 & 5 & 1 & 15 \\
2 & 3 & 0 & 1 & 3
\end{bmatrix}
\rightarrow \text{Interchange 1st and 2nd rows}
$$

$$
\longrightarrow
\begin{bmatrix}
1 & 2 & 1 & 0 & 4 \\
0 & 1 & 2 & 1 & 3 \\
3 & 7 & 5 & 1 & 15 \\
2 & 3 & 0 & 1 & 3
\end{bmatrix}
\rightarrow
\begin{array}{l}\text{Add multiple of 1st row by (-3) to 3rd row;}\\ \text{and add multiple of 1st row by (-2) to 4th row}\end{array}
$$

$$
\longrightarrow
\begin{bmatrix}
1 & 2 & 1 & 0 & 4 \\
0 & 1 & 2 & 1 & 3 \\
0 & 1 & 2 & 1 & 3 \\
0 & -1 & -2 & 1 & -5
\end{bmatrix}
\rightarrow
\begin{array}{l}\text{Add multiple of 2nd row by (-1) to 3rd row;}\\ \text{and add 2nd row to 4th row}\end{array}
$$

$$
\longrightarrow
\begin{bmatrix}
1 & 2 & 1 & 0 & 4 \\
0 & 1 & 2 & 1 & 3 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 2 & -2
\end{bmatrix}
\rightarrow
\begin{array}{l}\text{Multiply 4th row by } -\frac{1}{2};\\ \text{and then interchange 3rd and 4th rows}\end{array}
$$

$$
\longrightarrow
\begin{bmatrix}
1 & 2 & 1 & 0 & 4 \\
0 & 1 & 2 & 1 & 3 \\
0 & 0 & 0 & 1 & -1 \\
0 & 0 & 0 & 0 & 0
\end{bmatrix}
= B,
$$

which is in row echelon form. Since there are 3 leading 1's in B, the column-rank of A is 3. In addition, because the first, second and fourth columns of B are the columns containing leading 1's, the first, second and fourth columns of A form a basis for the column space of A; that is

$$
\begin{bmatrix} 0 \\ 2 \\ 3 \\ 2 \end{bmatrix}, \quad
\begin{bmatrix} 1 \\ 4 \\ 7 \\ 3 \end{bmatrix}, \quad
\begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \end{bmatrix}
$$

constitute a basis for the column space of A. ∎

Similarly, the row space of an $m \times n$ matrix $A \in \mathbb{C}^{m,n}$ is defined by the vector space spanned by the row vectors

$$
[a_{j1} \ \ldots \ a_{jn}], \ j = 1, \ldots, m
$$

of A. The "row-rank" of A is defined by the dimension of its row space. In other words, the **row-rank** of A is precisely the number of linearly independent rows of A. As shown in the following theorem, the row-rank and column-rank of any matrix are the same. Hence, the **rank** of a matrix A is defined by

$$
\text{rank}(A) := \text{row-rank}(A) = \text{column-rank}(A).
$$

Theorem 2 | **row-rank = column-rank** | *The row-rank and column-rank of any matrix are the same.*

Proof Suppose A is an $m \times n$ matrix and B is a row echelon form of A. Thus $B = EA$, where E is a nonsingular matrix. Therefore, every row of B is a linear combination of the rows of A. Hence,

$$\text{span}\{\text{rows of } B\} \subset \text{span}\{\text{rows of } A\},$$

which implies row-rank$(B) \leq$ row-rank(A). Similarly, since $A = E^{-1}B$, we also have row-rank$(A) \leq$ row-rank(B). Hence

$$\text{row-rank}(B) = \text{row-rank}(A).$$

On the other hand, it is obvious that being in row echelon form, the nonzero rows of B constitute a basis for the row space of B. Hence,

$$\text{row-rank}(B) = \# \text{ of nonzero rows of } B = \# \text{ of leading 1's in } B.$$

This, together with Theorem 1, leads to

$$\text{column-rank}(A) = \text{row-rank}(A) = \# \text{ of leading 1's in } B.$$

This completes the proof of the theorem.   ∎

Clearly for an $m \times n$ matrix A, we have rank$(A) \leq \min\{m, n\}$, rank$(A) =$ rank(A^T) = rank(A^*), and that a nonsingular linear transformation does not alter the rank of a matrix (see Exercise 16). Furthermore, for two matrices B, C,

$$\text{rank}(B + C) \leq \text{rank}(B) + \text{rank}(C), \tag{2.1.9}$$

$$\text{rank}(BC) \leq \min\{\text{rank}(B), \text{rank}(C)\}. \tag{2.1.10}$$

                                                                            ∎

Next, we recall that the **null space** of an $m \times n$ matrix A is defined by

$$\mathbb{N}_0(A) := \{\mathbf{x} \in \mathbb{C}^n : A\mathbf{x} = \mathbf{0}\}.$$

The dimension of the null space $\mathbb{N}_0(A)$ is called **nullity** of A, denoted by $\nu(A)$.

Observe that if B is a row echelon form or the reduced row echelon form of A, then $\mathbb{N}_0(A) = \mathbb{N}_0(B)$ since the two linear systems $A\mathbf{x} = \mathbf{0}$ and $B\mathbf{x} = \mathbf{0}$ have the same solution space. If $B\mathbf{x} = \mathbf{0}$ has the trivial solution only, then $\mathbb{N}_0(A) = \{\mathbf{0}\}$. In this case, $\mathbb{N}_0(A)$ has no basis and $\nu(A) = 0$. If $B\mathbf{x} = \mathbf{0}$ has a nontrivial solution, then from $B\mathbf{x} = \mathbf{0}$, it is not difficult to find a basis for $\mathbb{N}_0(B)$ (and hence for $\mathbb{N}_0(A)$ also). The dimensionof $\mathbb{N}_0(B)$ is equal to the number of the free parameters in the

solutions for $B\mathbf{x} = \mathbf{0}$. For example, for B given in (2.1.7), $x_{r+1}, \ldots, x_n$ are the free parameters in the solutions. Hence, in this case the nullity of B is $n - r$. Observe for an $m \times n$ matrix B in row echelon form or reduced row echelon form, the number of the free parameters in solutions for $B\mathbf{x} = \mathbf{0}$ is equal to

the number of columns of B *minus* the number of leading 1's in B.

Hence, we obtain the following formula on the rank and nullity of an arbitrary matrix studied in an elementary course in Linear Algebra or Matrix Theory.

Theorem 3 | **Rank and nullity formula** | *Let A be an $m \times n$ matrix. Then*

$$\mathrm{rank}(A) + \nu(A) = n. \tag{2.1.11}$$

Example 4 Let A be the matrix in Example 3. Find a basis for $\mathbb{N}_0(A)$.

Solution Let B be the matrix in row echelon form in Example 3. To find a basis for $\mathbb{N}_0(A)$, we only need find a basis for $\mathbb{N}_0(B)$. Since the solution of the linear system $B\mathbf{x} = \mathbf{0}$ is given by

$$\begin{cases} x_1 = 3x_3 + 4x_5, \\ x_2 = -2x_3 - 4x_5, \\ x_4 = x_5, \end{cases}$$

with two free parameters, x_3 and x_5, the dimension of the null space of A is 2, namely $\nu(A) = 2$. The general solution of $A\mathbf{x} = \mathbf{0}$ can be written as

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 3x_3 + 4x_5 \\ -2x_3 - 4x_5 \\ x_3 \\ x_5 \\ x_5 \end{bmatrix} = x_3 \begin{bmatrix} 3 \\ -2 \\ 1 \\ 0 \\ 0 \end{bmatrix} + x_5 \begin{bmatrix} 4 \\ -4 \\ 0 \\ 1 \\ 1 \end{bmatrix}.$$

Thus,

$$[3, -2, 1, 0, 0]^T, \quad [4, -4, 0, 1, 1]^T$$

form a basis of $\mathbb{N}_0(A)$.

Observe that B has 3 nonzero rows. Thus, $\mathrm{rank}(A) = \mathrm{rank}(B) = 3$, so that

$$\mathrm{rank}(A) + \nu(A) = 3 + 2 = 5,$$

which satisfies the formula (2.1.11) with $n = 5$. ∎

Example 5 Compute the rank and nullity of the matrix

$$A = \begin{bmatrix} 2 & 1 & 4 \\ -1 & -1 & -1 \\ 1 & 2 & -1 \\ 3 & 4 & 1 \end{bmatrix}.$$

Solution Since A has more rows than columns, we perform elementary row operations on A^T as follows to reduce the number of elementary row operations:

$$A^T = \begin{bmatrix} 2 & -1 & 1 & 3 \\ 1 & -1 & 2 & 4 \\ 4 & -1 & -1 & 1 \end{bmatrix} \rightarrow \text{ Interchange 1st and 2nd rows}$$

$$\longrightarrow \begin{bmatrix} 1 & -1 & 2 & 4 \\ 2 & -1 & 1 & 3 \\ 4 & -1 & -1 & 1 \end{bmatrix} \rightarrow \begin{array}{l} \text{Add (-2)-multiple of 1st row to 2nd row} \\ \text{and add (-4)-multiple of 1st row to 3rd row} \end{array}$$

$$\longrightarrow \begin{bmatrix} 1 & -1 & 2 & 4 \\ 0 & 1 & -3 & -5 \\ 0 & 3 & -9 & -15 \end{bmatrix} \rightarrow \text{ Add (-3)-multiple of 2nd row to 3rd row}$$

$$\longrightarrow \begin{bmatrix} 1 & -1 & 2 & 4 \\ 0 & 1 & -3 & -5 \\ 0 & 0 & 0 & 0 \end{bmatrix} = B,$$

which is in row echelon form. Thus, we obtain

$$\text{rank}(A) = \text{column-rank}(A) = \text{row-rank}(A^T) = 2.$$

In addition, since the dimension of A is 4×3 (with $n = 3$), it follows from (2.1.11) that the nullity of A is

$$\nu(A) = 3 - 2 = 1.$$

∎

We conclude this section by showing that BB^*, B^*B and B have the same rank.

Theorem 4 $\boxed{\textbf{B}, \textbf{BB}^*, \text{ and } \textbf{B}^*\textbf{B} \text{ have the same rank}}$ *For any matrix B,*

$$\text{rank}(BB^*) = \text{rank}(B^*B) = \text{rank}(B).$$

Proof For an arbitrary matrix $B \in \mathbb{C}^{m,n}$, set $C = B^*B$, and observe that

$$\mathbb{N}_0(C) = \mathbb{N}_0(B). \tag{2.1.12}$$

Indeed, if $\mathbf{x} \in \mathbb{N}_0(B)$, then $C\mathbf{x} = B^*(B\mathbf{x}) = B^*\mathbf{0} = \mathbf{0}$, so that $\mathbf{x} \in \mathbb{N}_0(C)$. On the other hand, for any $\mathbf{x} \in \mathbb{N}_0(C)$, we have

$$0 = \langle \mathbf{x}, \mathbf{0} \rangle = \langle \mathbf{x}, C\mathbf{x} \rangle = \langle \mathbf{x}, B^*B\mathbf{x} \rangle$$
$$= \langle B\mathbf{x}, B\mathbf{x} \rangle = ||B\mathbf{x}||^2,$$

so that $B\mathbf{x} = \mathbf{0}$; that is, $\mathbf{x} \in \mathbb{N}_0(B)$.

Now, since C is an $n \times n$ matrix, it follows from (2.1.11) and (2.1.12) that

$$\text{rank}(C) = n - \nu(C) = n - \nu(B) = \text{rank}(B),$$

or $\text{rank}(B^*B) = \text{rank}(B)$. Finally, since $BB^* = (B^*B)^*$, we may conclude that $\text{rank}(BB^*) = \text{rank}(B^*B)$, which agrees with $\text{rank}(B)$. ∎

Exercises

Exercise 1 Consider the matrices

$$A = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 2 & 0 \\ 1 & 1 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}.$$

Which of the following nine multiplications AB, AC, BA, BC, CA, CB, A^2, B^2, C^2 do not make sense? Why? Compute those that are well defined.

Exercise 2 Let A, B, C be the matrices given in Exercise 1. Which of the following matrix operations do not make sense? Why? Compute those that are well defined.

(a) $AC + 2B$.
(b) $AC - A$.
(c) $BA + 3C$.
(d) $BA + 2A$.

Exercise 3 Let B, C, D, E be 2×2 matrices. Show that

$$\begin{bmatrix} B & O \\ O & C \end{bmatrix} \begin{bmatrix} D & O \\ O & E \end{bmatrix} = \begin{bmatrix} BD & O \\ O & CE \end{bmatrix},$$

where O is the zero matrix (with appropriate size).

Exercise 4 Repeat Exercise 3, where B, C, D and E are 2×3, 2×2, 3×2 and 2×3 matrices, respectively.

Exercise 5 Use the block matrix multiplication formula in Exercise 3 to compute the matrix product $\text{diag}\{B, C\}\text{diag}\{D, E\}$, where

$$B = \begin{bmatrix} 1 & 1 \\ -1 & 0 \end{bmatrix}, \quad C = \begin{bmatrix} 0.5 & 3 \\ -1.5 & 1 \end{bmatrix}, \quad D = \begin{bmatrix} 1 & -1 \\ 1 & 2 \end{bmatrix}, \quad E = \begin{bmatrix} 2 & 0 \\ 0 & -1 \end{bmatrix}.$$

Exercise 6 Use the block matrix multiplication formula in Exercise 4 to calculate the matrix product $\text{diag}\{B, C\}\text{diag}\{D, E\}$, where

$$B = \begin{bmatrix} 1 & 2 & 1 \\ 1 & 2 & -1 \end{bmatrix}, \quad C = \begin{bmatrix} -1 & 1 \\ 0 & 1 \end{bmatrix}, \quad D = \begin{bmatrix} 3 & 1 \\ -1 & 0 \\ -2 & -1 \end{bmatrix}, \quad E = \begin{bmatrix} 0 & 1 & 3 \\ 1 & -1 & 2 \end{bmatrix}.$$

Exercise 7 Suppose that the vectors $\mathbf{a}_j$ and $\mathbf{b}_j$, $1 \le j \le r$ in $\mathbb{C}^n$ satisfy $\mathbf{a}_j = E\mathbf{b}_j$, where E is a nonsingular matrix. Show that $\mathbf{a}_1, \ldots, \mathbf{a}_r$ are linearly independent if and only if $\mathbf{b}_1, \ldots, \mathbf{b}_r$ are linearly independent.

Exercise 8 Determine the vectors from

$$\mathbf{v}_1 = \begin{bmatrix} 1 \\ 2 \\ 2 \\ 3 \\ 2 \end{bmatrix}, \mathbf{v}_2 = \begin{bmatrix} 2 \\ 4 \\ 4 \\ 6 \\ 4 \end{bmatrix}, \mathbf{v}_3 = \begin{bmatrix} 1 \\ 2 \\ 2 \\ 3 \\ 1 \end{bmatrix}, \mathbf{v}_4 = \begin{bmatrix} 2 \\ 5 \\ 4 \\ 7 \\ 3 \end{bmatrix}, \mathbf{v}_5 = \begin{bmatrix} 1 \\ 4 \\ 2 \\ 5 \\ 3 \end{bmatrix}$$

that form a basis of the vector space span$\{\mathbf{v}_1, \ldots, \mathbf{v}_5\}$.

Exercise 9 Repeat Exercise 8 with

$$\mathbf{v}_1 = (0, \ 2, \ 3, \ 5, \ 6), \quad \mathbf{v}_2 = (2, \ 4, \ -6, \ 0, \ -2),$$
$$\mathbf{v}_3 = (3, \ 2, \ 1, \ -2, \ 1), \quad \mathbf{v}_4 = (2, \ -2, \ 1, \ 3, \ 4).$$

Exercise 10 Perform elementary row operations on each of the following matrices to (i) compute its row-rank and (ii) find a basis for its column space. Then verify that the column-rank and row-rank for each of these matrices are the same.

(a) $A_1 = \begin{bmatrix} -3 & 2 & -1 & 0 \\ 1 & 0 & 2 & 1 \\ 0 & 2 & 5 & 3 \end{bmatrix}$.

(b) $A_2 = \begin{bmatrix} 1 & -1 & 2 \\ -1 & 2 & 1 \\ 0 & 1 & 3 \\ 1 & 0 & 5 \end{bmatrix}$.

(c) $A_3 = \begin{bmatrix} 0 & 2 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & 5 & 4 & 0 \\ 4 & 2 & 1 & 3 \end{bmatrix}$.

Exercise 11 For each of the matrices A_1, A_2, A_3 in Exercise 10, perform elementary row operations to the transpose of the matrix to (i) compute its column-rank and (ii) find a basis for its row space. Then verify that the column-rank and row-rank for each of A_1, A_2, A_3 are the same.

Exercise 12 Show that for any matrix A, rank$(A) = $ rank(A^T).

Exercise 13 Suppose A and B are two matrices of sizes $m \times n$ and $n \times m$, respectively. Determine whether $\mathrm{rank}(AB) = \mathrm{rank}(BA)$ always holds or not. If yes, give a proof for this formula. Otherwise, give a counter-example.

Exercise 14 Show that for two matrices B and C of the same size, the inequality (2.1.9) holds.

Exercise 15 Show that for two matrices B and C with appropriate sizes, the inequality (2.1.10) holds.

Exercise 16 Show that a nonsingular matrix transformation of a rectangular matrix B does not change its rank; that is, $\mathrm{rank}(B) = \mathrm{rank}(AB) = \mathrm{rank}(BC)$, where A and C are nonsingular matrices.

Exercise 17 Let A_1, A_2, A_3 be the matrices given in Exercise 10. Answer the following questions.

(a) Is $\mathrm{rank}(A_1 A_2) = \mathrm{rank}(A_2 A_1)$?
(b) Is $\mathrm{rank}(A_1 A_3^T) = \mathrm{rank}(A_1)$?
(c) Is $\mathrm{rank}(A_1 A_3^T) = \mathrm{rank}(A_3)$?
(d) Is $\mathrm{rank}(A_2^T A_3^T) \leq \min\{\mathrm{rank}(A_2), \mathrm{rank}(A_3)\}$?

2.2 Linear Transformations

Multiplication of an $m \times n$ matrix A to a column vector $\mathbf{x} \in \mathbb{C}^n$ results in a column vector $\mathbf{z} \in \mathbb{C}^m$. Hence, the matrix A can be considered as a mapping (to be called a transformation in this book) from the Euclidean space $\mathbb{C}^n$ to the Euclidean space $\mathbb{C}^m$. This mapping has the important property that

$$A(a\mathbf{x} + b\mathbf{y}) = aA\mathbf{x} + bA\mathbf{y}$$

for all vectors $\mathbf{x}, \mathbf{y} \in \mathbb{C}^n$ and all scalars $a, b \in \mathbb{C}$. To be called the linearity property of the matrix A, this concept extends to transformations defined on finite or infinite-dimensional vector spaces, including fairly general differential and integral operators on certain appropriate subspaces of the function spaces $\widetilde{L}_p(J)$ as studied in the previous chapter.

Definition 1 $\boxed{\textbf{Linear transformations}}$ *Let $\mathbb{V}$ and $\mathbb{W}$ be two vector spaces over some scalar field $\mathbb{F}$, such as $\mathbb{C}$ or $\mathbb{R}$. A transformation T from $\mathbb{V}$ to $\mathbb{W}$ is said to be linear, if it satisfies:*

$$T(a\mathbf{x} + b\mathbf{y}) = aT\mathbf{x} + bT\mathbf{y} \tag{2.2.1}$$

for all $\mathbf{x}, \mathbf{y} \in \mathbb{V}$ and $a, b \in \mathbb{F}$.

It follows from (2.2.1) that a linear transformation T satisfies

$$T(\mathbf{x} + \mathbf{y}) = T\mathbf{x} + T\mathbf{y}; \qquad (2.2.2)$$
$$T(a\mathbf{x}) = aT\mathbf{x} \qquad (2.2.3)$$

for all $\mathbf{x}, \mathbf{y} \in \mathbb{V}$ and $a \in \mathbb{F}$. Conversely, (2.2.1) follows from (2.2.2) and (2.2.3) as well (see Exercise 1).

Example 1 As mentioned above, any matrix $A \in \mathbb{C}^{m,n}$ is a linear transformation from $\mathbb{C}^n$ to $\mathbb{C}^m$. This statement remains valid if $\mathbb{C}^n$, $\mathbb{C}$ are replaced by $\mathbb{R}^n$, $\mathbb{R}$, respectively.

Validity of the statement in this example follows from the rule of matrix multiplication, with B being an $n \times 1$ matrix; and scalar multiplication as defined in Sect. 1.1 of Chap. 1. Details will be left as an exercise (see Exercise 2). ∎

Observe that the scalar field $\mathbb{F}$ can be considered as a vector space over $\mathbb{F}$ itself. For instance, in the above example, the matrix $A \in \mathbb{C}^{1 \times n}$ (with $m = 1$), is a linear transformation from $\mathbb{V} = \mathbb{C}^n$ to the vector space $\mathbb{W} = \mathbb{C}$, which is the scalar field $\mathbb{C}$. This matrix transformation, from a vector space to a scalar field, is called a linear functional. More generally, we have the following.

Definition 2 | **Linear functionals**, **linear operators** | *A linear transformation from a vector space $\mathbb{V}$ (over some scalar field $\mathbb{F}$, such as $\mathbb{C}$ or $\mathbb{R}$) to $\mathbb{F}$ is called a linear functional on $\mathbb{V}$. A linear transformation T from a vector space $\mathbb{V}$ to itself is called a linear operator on $\mathbb{V}$.*

If the vector spaces $\mathbb{V}$ and $\mathbb{W}$ are both normed (linear) spaces, then a linear transformation T from $\mathbb{V}$ to $\mathbb{W}$ is said to be bounded, provided that the norm of $T\mathbf{x}$ is uniformly bounded for all unit vectors $\mathbf{x} \in \mathbb{V}$. For bounded linear transformations, we introduce the notion of operator norm, as follows.

Definition 3 | **Operator norm** | *Let $\mathbb{V}$ and $\mathbb{W}$ be two normed spaces (over some scalar field $\mathbb{F}$, such as $\mathbb{C}$ or $\mathbb{R}$) with norms denoted by $\|\cdot\|_{\mathbb{V}}$ and $\|\cdot\|_{\mathbb{W}}$, respectively. The norm of a linear transformation T from $\mathbb{V}$ to $\mathbb{W}$ is defined by*

$$\|\|T\|\|_{\mathbb{V} \to \mathbb{W}} = \sup \left\{ \frac{\|T\mathbf{x}\|_{\mathbb{W}}}{\|\mathbf{x}\|_{\mathbb{V}}} : \mathbf{0} \neq \mathbf{x} \in \mathbb{V} \right\}. \qquad (2.2.4)$$

If $\|\|T\|\|_{\mathbb{V} \to \mathbb{W}}$ is finite, then the transformation T is said to be a bounded transformation.

For convenience, if $\mathbb{W} = \mathbb{V}$ (that is, for linear operators T on $\mathbb{V}$), we will adopt the abbreviated notation

$$\|\|T\|\|_{\mathbb{V}} = \|\|T\|\|_{\mathbb{V} \to \mathbb{V}}. \qquad (2.2.5)$$

Example 2 Let $\mathbb{V} = \widetilde{L}_1(J)$, where J is an interval on $\mathbb{R}$. Then

$$Tf = \int_J f \tag{2.2.6}$$

is a bounded linear functional on $\mathbb{V}$.

Example 3 Let $\mathbb{V} = C(J)$ be the vector space of continuous functions on the interval J over $\mathbb{C}$ or $\mathbb{R}$, as introduced in Sect. 1.2, and let $b \in J$. Then

$$Tf = f(b), \quad f \in C(J), \tag{2.2.7}$$

is a bounded linear functional on $C(J)$.

Example 4 Let $\mathbb{V}$ be an inner-product space with inner product $\langle \cdot, \cdot \rangle$, and let $\mathbf{x}_0 \in \mathbb{V}$. Then

$$T\mathbf{x} = \langle \mathbf{x}, \mathbf{x}_0 \rangle, \quad \mathbf{x} \in \mathbb{V},$$

is a bounded linear functional from $\mathbb{V}$ to $\mathbb{C}$ with $\|T\|_{\mathbb{V} \to \mathbb{C}} = \|\mathbf{x}_0\|$.

Verifications of Examples 2–4 are left as exercises (see Exercises 8–10).

The converse of Example 4 is also valid by employing a pair of dual bases, and in particular, an orthonormal basis. Recall from Definition 5 on p.39, that two bases $\{\mathbf{v}_k, k = 1, 2, \ldots\}$ and $\{\widetilde{\mathbf{v}}_k, k = 1, 2, \ldots\}$ of an inner-product space $\mathbb{V}$ are said to constitute a dual pair, if

$$\langle \mathbf{v}_k, \widetilde{\mathbf{v}}_j \rangle = \delta_{j-k}, \quad j, k = 1, 2, \ldots,$$

where $\langle \cdot, \cdot \rangle$ is the inner product for the space $\mathbb{V}$.

Theorem 1 *Let $\mathbb{V}$ be a complete inner-product space with dual bases $\{\mathbf{v}_k\}$ and $\{\widetilde{\mathbf{v}}_k\}$, and let T be a bounded linear functional on $\mathbb{V}$, such that $\mathbf{x}_T$, defined by*

$$\mathbf{x}_T = \sum_j \overline{(T\mathbf{v}_j)}\widetilde{\mathbf{v}}_j, \tag{2.2.8}$$

is in $\mathbb{V}$. Then $T: \mathbb{V} \to \mathbb{F}$ can be formulated as:

$$T\mathbf{x} = \langle \mathbf{x}, \mathbf{x}_T \rangle \text{ for all } \mathbf{x} \in \mathbb{V}. \tag{2.2.9}$$

Furthermore, $\mathbf{x}_T$ in (2.2.8), called the representer of T, is unique.

Proof To derive the representation (2.2.9) for each $\mathbf{x} \in \mathbb{V}$, write

$$\mathbf{x} = \sum_k c_k \mathbf{v}_k,$$

where the series converges in the sense of (1.4.4) on p.38, with $\| \cdot \| = \sqrt{\langle \cdot, \cdot \rangle}$. Then by linearity and taking the limit, we have (see Exercise 12)

$$T\mathbf{x} = \sum_k c_k T\mathbf{v}_k. \tag{2.2.10}$$

On the other hand, again by linearity and taking limit, we also have, by the definition of $\mathbf{x}_T$ given in (2.2.8),

$$
\begin{aligned}
\langle \mathbf{x}, \mathbf{x}_T \rangle &= \sum_k \left\langle c_k \mathbf{v}_k, \sum_j \overline{T\mathbf{v}_j} \widetilde{\mathbf{v}}_j \right\rangle \\
&= \sum_k c_k \sum_j T\mathbf{v}_j \langle \mathbf{v}_k, \widetilde{\mathbf{v}}_j \rangle \\
&= \sum_k c_k \sum_j T\mathbf{v}_j \delta_{k-j} = \sum_k c_k T\mathbf{v}_k.
\end{aligned}
$$

Hence, it follows from (2.2.10) that (2.2.9) is established.

To prove that $\mathbf{x}_T$ in (2.2.9) is unique, let $\mathbf{y}_0 \in \mathbb{V}$ be another represeter of T. Then

$$T\mathbf{x} = \langle \mathbf{x}, \mathbf{x}_T \rangle = \langle \mathbf{x}, \mathbf{y}_0 \rangle, \quad \text{for all } \mathbf{x} \in \mathbb{V}$$

so that

$$\langle \mathbf{x}, \mathbf{x}_T - \mathbf{y}_0 \rangle = 0.$$

By choosing $\mathbf{x} = \mathbf{x}_T - \mathbf{y}_0$, we have

$$\| \mathbf{x}_T - \mathbf{y}_0 \|^2 = 0,$$

or $\mathbf{y}_0 = \mathbf{x}_T$.   ■

Remark 1 We remark that $\mathbf{x}_T$, as defined in (2.2.8), is always in $\mathbb{V}$. Refer to Exercise 11 for the proof in the case $\{\mathbf{v}_k\} = \{\widetilde{\mathbf{v}}_k\}$ (that is, $\{\mathbf{v}_k\}$ is an orthonormal basis of $\mathbb{V}$). However, the proof for the general case of dual bases is beyond the scope of this book.   ■

Throughout the entire book, the inner product $\langle \cdot, \cdot \rangle$ for the Euclidean space $\mathbb{C}^n$ or $\mathbb{R}^n$ is always the ordinary dot product, so that the corresponding norm $\| \cdot \|$ of $\mathbf{z}$ in $\mathbb{C}^n$ or $\mathbb{R}^n$ is its Euclidean norm (or length of the vector $\mathbf{z}$):

$$\| \mathbf{z} \| = (|z_1|^2 + \cdots + |z_n|^2)^{\frac{1}{2}}.$$

Example 5 Let $S = \{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ and $\widetilde{S} = \{\widetilde{\mathbf{v}}_1, \widetilde{\mathbf{v}}_2, \widetilde{\mathbf{v}}_3\}$ be dual bases of $\mathbb{R}^3$ as given in Example 4 on p.40. Consider the linear functional T, defined by

$$Tx = \sum_{j=1}^{3} x_j, \qquad (2.2.11)$$

where $\mathbf{x} = (x_1, x_2, x_3) \in \mathbb{R}^3$. Write down the representer $\mathbf{x}_T$ of T with dual bases S and $\widetilde{S}$, and verify that

$$Tx = \langle \mathbf{x}, \mathbf{x}_T \rangle$$

for all $\mathbf{x} \in \mathbb{R}^3$.

Solution In view of (2.2.8) in Theorem 1, the required representer is given by

$$
\begin{aligned}
\mathbf{x}_T &= \sum_{j=1}^{3} (T\mathbf{v}_j)\widetilde{\mathbf{v}}_j \\
&= (T\mathbf{v}_1)\widetilde{\mathbf{v}}_1 + (T\mathbf{v}_2)\widetilde{\mathbf{v}}_2 + (T\mathbf{v}_3)\widetilde{\mathbf{v}}_3 \\
&= 2\,\widetilde{\mathbf{v}}_1 + 0\,\widetilde{\mathbf{v}}_2 + 3\,\widetilde{\mathbf{v}}_3 \\
&= 2(-1, 2, -1) + (0, 0, 0) + 3(1, -1, 1) \\
&= (1, 1, 1).
\end{aligned}
$$

Hence, for each $\mathbf{x} = (x_1, x_2, x_3) \in \mathbb{R}^3$,

$$\langle \mathbf{x}, \mathbf{x}_T \rangle = x_1(1) + x_2(1) + x_3(1) = \sum_{j=1}^{3} x_j.$$

Thus, $\langle \mathbf{x}, \mathbf{x}_T \rangle$ is exactly Tx in (2.2.11).

On the other hand, if we use the standard basis

$$\mathbf{e}_1 = (1, 0, 0), \ \ \mathbf{e}_2 = (0, 1, 0), \ \ \mathbf{e}_3 = (0, 0, 1),$$

of $\mathbb{R}^3$, then by applying (2.2.8) with $\mathbf{v}_j = \widetilde{\mathbf{v}}_j = \mathbf{e}_j$, we again have

$$
\begin{aligned}
\mathbf{x}_T &= \sum_{j=1}^{3} (T\mathbf{e}_j)\mathbf{e}_j \\
&= (T\mathbf{e}_1)\mathbf{e}_1 + (T\mathbf{e}_2)\mathbf{e}_2 + (T\mathbf{e}_3)\mathbf{e}_3 \\
&= \mathbf{e}_1 + \mathbf{e}_2 + \mathbf{e}_3 = (1, 1, 1),
\end{aligned}
$$

which agrees with the above result. This example illustrates the independence of the representer $\mathbf{x}_T$ in terms of the choice of dual bases in (2.2.8). ∎

Example 6 For $n \geq 1$, show that any square matrix $A \in \mathbb{C}^{n,n}$ is a bounded linear operator on the vector space $\mathbb{C}^n$.

In view of the linearity property, it is sufficient to show that $\|A\mathbf{x}\|$ is uniformly bounded for all $\mathbf{x} \in \mathbb{C}^n$ with $\|\mathbf{x}\| = 1$ (see Exercise 6). ∎

We next introduce the notion of adjoints of linear operators, but would like to point out in advance that this concept differs from the definition of (matrix) adjoints of square matrices in an elementary course of Linear Algebra or Matrix Theory (see Remark 2 on p.85).

Theorem 2 | **Adjoint of linear operators** | *Let $\mathbb{V}$ be a complete inner-product space over the scalar field $\mathbb{C}$ or $\mathbb{R}$, such that certain dual bases of $\mathbb{V}$ exist. Then corresponding to any bounded linear operator T on $\mathbb{V}$, there exists a linear operator T^* defined on $\mathbb{V}$, called the adjoint of T, that satisfies the property*

$$\langle T\mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{x}, T^*\mathbf{y} \rangle, \quad \text{for all } \mathbf{x}, \mathbf{y} \in \mathbb{V}. \tag{2.2.12}$$

Furthermore, T^* is uniquely determined by (2.2.12).

Proof In view of Theorem 1 and Remark 1, we observe that for any fixed $\mathbf{y} \in \mathbb{V}$, the linear functional $L = L_\mathbf{y}$, defined by

$$L_\mathbf{y}\mathbf{x} = \langle T\mathbf{x}, \mathbf{y} \rangle, \tag{2.2.13}$$

is bounded, and hence, has a unique representer $\mathbf{x}_{L_\mathbf{y}}$, in that $L_\mathbf{y}\mathbf{x} = \langle \mathbf{x}, \mathbf{x}_{L_\mathbf{y}} \rangle$ for all $\mathbf{x} \in \mathbb{V}$. Since $\mathbf{x}_{L_\mathbf{y}}$ is a function of $\mathbf{y}$, we may write $\mathbf{x}_{L_\mathbf{y}} = F(\mathbf{y})$, so that

$$\langle T\mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{x}, F(\mathbf{y}) \rangle \tag{2.2.14}$$

for all $\mathbf{x} \in \mathbb{V}$, where (2.2.14) holds for all $\mathbf{y} \in \mathbb{V}$. Next, let us verify that F is a linear operator on $\mathbb{V}$. From its definition, F is a transformation from $\mathbb{V}$ into itself. That F is a linear operator on $\mathbb{V}$ follows from the linearity of the inner product. Indeed, for any fixed $\mathbf{y}, \mathbf{z} \in \mathbb{V}$ and fixed $a, b \in \mathbb{C} = \mathbb{F}$, it follows from (2.2.14) that, for all $\mathbf{x} \in \mathbb{V}$,

$$\begin{aligned}
\langle \mathbf{x}, F(a\mathbf{y} + b\mathbf{z}) \rangle &= \langle T\mathbf{x}, a\mathbf{y} + b\mathbf{z} \rangle \\
&= \langle T\mathbf{x}, a\mathbf{y} \rangle + \langle T\mathbf{x}, b\mathbf{z} \rangle \\
&= \overline{a}\langle T\mathbf{x}, \mathbf{y} \rangle + \overline{b}\langle T\mathbf{x}, \mathbf{z} \rangle \\
&= \overline{a}\langle \mathbf{x}, F(\mathbf{y}) \rangle + \overline{b}\langle \mathbf{x}, F(\mathbf{z}) \rangle \\
&= \langle \mathbf{x}, aF(\mathbf{y}) \rangle + \langle \mathbf{x}, bF(z) \rangle,
\end{aligned}$$

so that $\langle \mathbf{x}, F(a\mathbf{y} + b\mathbf{z}) - (aF(\mathbf{y}) + bF(\mathbf{z})) \rangle = 0$ for all $\mathbf{x} \in \mathbb{V}$. By setting $\mathbf{x} = F(a\mathbf{y} + b\mathbf{z}) - (aF(\mathbf{y}) + bF(\mathbf{z}))$, we have

$$\|F(a\mathbf{y} + b\mathbf{z}) - (aF(\mathbf{y}) + bF(\mathbf{z}))\|^2 = 0;$$

that is, $F(a\mathbf{y} + b\mathbf{z}) = aF(\mathbf{y}) + bF(\mathbf{z})$. Hence, by setting $T^* = F$, we have derived (2.2.12). Furthermore, since the representer in Theorem 1 is unique, $T^* = F$ is unique. ∎

Example 7 Consider an $n \times n$ matrix $A \in \mathbb{C}^{n,n}$ as a linear operator on the vector space $\mathbb{V} = \mathbb{C}^n$. Determine the adjoint A^* of A.

Solution For any $\mathbf{x}, \mathbf{y} \in \mathbb{C}^n$, consider $\mathbf{x}$ and $\mathbf{y}$ as column vectors:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

so that $\mathbf{x}^T = [x_1, \ldots, x_n]$ and $\mathbf{y}^T = [y_1, \ldots, y_n]$ are row vectors (where the superscript T denotes, as usual, the transpose of the matrix). Hence, it follows from the definition of the inner product for $\mathbb{C}^n$ that

$$\begin{aligned} \langle A\mathbf{x}, \mathbf{y} \rangle &= (A\mathbf{x})^T \overline{\mathbf{y}} = \mathbf{x}^T A^T \overline{\mathbf{y}} \\ &= \mathbf{x}^T \overline{\left((\overline{A})^T \mathbf{y} \right)} = \langle \mathbf{x}, (\overline{A})^T \mathbf{y} \rangle, \end{aligned}$$

so that from the uniqueness of the adjoint A^* of A, we have

$$A^* = (\overline{A})^T. \tag{2.2.15}$$

In other words, the adjoint of A is the transpose-conjugate A^* of A defined in (2.1.5) on p.68. ∎

As pointed out previously, the operator adjoint is different from the matrix adjoint being studied in an elementary course of matrix theory or linear algebra, as follows.

Remark 2 ⎡Matrix adjoints⎤ In Matrix Theory, the adjoint (which we call matrix adjoint) of a square matrix $A = [a_{j,k}]_{1 \le j,k \le n}$ is the matrix

$$\left([A_{j,k}]_{1 \le j,k \le n} \right)^T,$$

where $A_{j,k}$ denotes the cofactor of $a_{j,k}$. Hence in general, the matrix adjoint of A is different from the operator adjoint $A^* = (\overline{A})^T$, when A is considered as an operator on the vector space $\mathbb{C}^n$. ∎

Definition 4 ⎡Self-adjoint operators⎤ *A linear operator T is said to be self-adjoint if $T^* = T$.*

Remark 3 ⎡Hermitian matrices⎤ Recall that in an elementary course on Linear Algebra or Matrix Theory, a square matrix $A \in \mathbb{C}^{n,n}$ is said to be Hermitian, if

$A = \left(\overline{A}\right)^T$. Thus, in view of (2.2.15), when considered as a linear operator, a square matrix $A \in \mathbb{C}^{n,n}$ is self-adjoint if and only if it is Hermitian. Henceforth, for linear operators T, which may not be square matrices, we say that T is Hermitian if it is self-adjoint. Clearly, if a matrix $A \in \mathbb{C}^{n \times n}$ is Hermitian, then all the diagonal entries of A are real. It will be shown in the next section that all eigenvalues of self-adjoint operators are real in general.   ∎

Example 8 Write down the (operator) adjoints $A_1^*, \ldots, A_4^*$ for the following matrices:

$$A_1 = \begin{bmatrix} 1 & 4 \\ 3 & 2 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 1 & -i \\ i & 2 \end{bmatrix},$$

$$A_3 = \begin{bmatrix} 1 & i \\ i & 2 \end{bmatrix}, \quad A_4 = \begin{bmatrix} 2 & 1-i \\ 1+i & 3 \end{bmatrix}.$$

Solution By applying (2.2.15), we have

$$A_1^* = \begin{bmatrix} 1 & 3 \\ 4 & 2 \end{bmatrix}, \quad A_2^* = \begin{bmatrix} 1 & -i \\ i & 2 \end{bmatrix},$$

and

$$A_3^* = \begin{bmatrix} 1 & -i \\ -i & 2 \end{bmatrix}, \quad A_4^* = \begin{bmatrix} 2 & 1-i \\ 1+i & 3 \end{bmatrix}.$$

Observe that A_3 and A_3^* are both symmetric, but they are not Hermitian. On the other hand, $A_2^* = A_2$ and $A_4^* = A_4$. Thus A_2 and A_4 are Hermitian, but they are not symmetric. Finally, A_1 and A_1^* are neither symmetric nor Hermitian.   ∎

<div align="center">

Exercises

</div>

Exercise 1 Show that (2.2.1) is equivalent to the totality of (2.2.2) and (2.2.3).

Exercise 2 Complete the solution of Example 1 by filling in more details.

Exercise 3 Recall that $\mathbb{R}$ can be considered as a vector space $\mathbb{V}$ over $\mathbb{R}$ itself. Hence, any function $f : \mathbb{R} \to \mathbb{R}$ (that is, $y = f(x) \in \mathbb{R}$ for all $x \in \mathbb{R}$) is a transformation of $\mathbb{V} = \mathbb{R}$ to $\mathbb{W} = \mathbb{R}$. Identify all linear transformations among the functions in the following by verifying the condition (2.2.1) in Definition 1. Justify your answers with reasoning.

(a) $f(x) = x^2$, for $x \in \mathbb{R}$.
(b) $g(x) = 2x$, for $x \in \mathbb{R}$.
(c) $h(x) = -5x$, for $x \in \mathbb{R}$.
(d) $k(x) = x + 2$, for $x \in \mathbb{R}$.

Exercise 4 Let $\mathbb{V} = \mathbb{C}^{m,n}$ be the set of $m \times n$ matrices with complex entries, with scalar multiplication, matrix addition, and matrix multiplication defined in Sect. 2.1.

In particular, for $V = \mathbb{C}^n$, $W = \mathbb{C}^m$, and $A \in \mathbb{C}^{m,n}$, $\mathbf{y} = A\mathbf{x}$ is a vector in $W = \mathbb{C}^m$ for all $\mathbf{x} \in V = \mathbb{C}^n$. Identify all linear transformations from $\mathbb{C}^2$ to $\mathbb{C}^3$ among the following operations $T_1, \ldots, T_4$, by verifying the condition (2.2.1) in Definition 1. Justify your answers with rigorous arguments.

(a) $T_1\mathbf{x} = \begin{bmatrix} -1 & 0 \\ 1 & 2 \\ 0 & 0 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$, for $\mathbf{x} \in \mathbb{C}^2$.

(b) $T_2\mathbf{x} = \begin{bmatrix} -1 & 0 \\ 0 & 0 \\ 1 & 2 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 1 & 0 \\ 2 & 1 \\ 0 & -1 \end{bmatrix} \mathbf{x}$, for $\mathbf{x} \in \mathbb{C}^2$.

(c) $T_3\mathbf{x} = 3 \begin{bmatrix} 1 & 0 \\ 0 & i \\ -i & 1 \end{bmatrix} \mathbf{x}$, for $\mathbf{x} \in \mathbb{C}^2$.

(d) $T_4\mathbf{x} = \begin{bmatrix} 1 & 0 \\ 2-i & 3i \\ -i & 4 \end{bmatrix} \mathbf{x}$, for $\mathbf{x} \in \mathbb{C}^2$.

Exercise 5 As a continuation of Exercise 4, identify all linear transformations from $\mathbb{C}^3$ to $\mathbb{C}^2$ among $S_1, \ldots, S_4$ by verifying (2.2.1) in Definition 1. Justify your answers with rigorous arguments.

(a) $S_1\mathbf{x} = \begin{bmatrix} -1 & 1 & 0 \\ 0 & 2 & 0 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, for $\mathbf{x} \in \mathbb{C}^3$.

(b) $S_2\mathbf{x} = \begin{bmatrix} i & -1 & 0 \\ 0 & 1 & i \end{bmatrix} \mathbf{x} - \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \mathbf{x}$, for $\mathbf{x} \in \mathbb{C}^3$.

(c) $S_3\mathbf{x} = 3 \begin{bmatrix} 1 & 0 & -i \\ 0 & i & 0 \end{bmatrix} \mathbf{x}$, for $\mathbf{x} \in \mathbb{C}^3$.

(d) $S_4\mathbf{x} = \begin{bmatrix} 1 & 2-i & i-1 \\ 0 & 3i & 4 \end{bmatrix} \mathbf{x}$, for $\mathbf{x} \in \mathbb{C}^3$.

Exercise 6 Verify that any square matrix A is a bounded linear operator when considered as a linear transformation.

Hint: Apply the Cauchy-Schwarz inequality to show that $\|\|A\|\|_{\mathbb{C}^n}$ is bounded by the summation of the Euclidean norms of the row vectors of A.

Exercise 7 Let $\mathbb{T}$ be the set of all linear transformations from a vector space V over some scalar field $\mathbb{F}$ to a vector space W over the same scalar field $\mathbb{F}$. For any $T, S \in \mathbb{T}$ and any $a \in \mathbb{F}$, define $T + S$ and aT by

$$(T + S)\mathbf{x} = T\mathbf{x} + S\mathbf{x}, \text{ for } \mathbf{x} \in V$$

and

$$(aT)\mathbf{x} = a(T\mathbf{x}), \text{ for } \mathbf{x} \in V.$$

Then $\mathbb{T}$ can be considered as a vector space (of linear transformations) over $\mathbb{F}$. Justify this statement by considering the set $\mathbb{C}^{m,n}$ of $m \times n$ matrices.

Exercise 8 Verify that the transformation T defined in (2.2.6) is a bounded linear functional on $\widetilde{L}_1(J)$, with $\|\|T\|\|_{\mathbb{V}\to\mathbb{C}} = 1$.

Exercise 9 Verify that the transformation T defined in (2.2.7) is a bounded linear functional on $C(J)$, with $\|\|T\|\|_{\mathbb{V}\to\mathbb{C}} = 1$.

Exercise 10 Verify that the transformation T defined in (2.2.9), where $\mathbf{x}_T$ is fixed, is a bounded linear functional on $\mathbb{V}$. This is the solution of Example 4 with $\mathbf{x}_0 = \mathbf{x}_T$.

   Hint: Analogous to Exercise 6, apply the Cauchy-Schwarz inequality to derive an upper bound of the operator norm of T.

Exercise 11 Let T be a bounded linear functional on a complete inner-product space $\mathbb{V}$, namely $|T\mathbf{x}| \leq c\|\mathbf{x}\|$ for all $\mathbf{x} \in \mathbb{V}$, where $c > 0$ is a constant. Also, let $\{\mathbf{v}_k\}_{k=1}^{\infty}$ be an orthonormal basis for $\mathbb{V}$.

(a) Show that $\|\sum_{k=1}^{N}\langle \mathbf{x}, \mathbf{v}_k\rangle \mathbf{v}_k\|_2 \leq \|\mathbf{x}\|$, for any $N \in \mathbb{N}$, $\mathbf{x} \in \mathbb{V}$.
(b) Apply (a) and the boundedness of T to show that $|\langle \mathbf{x}, \sum_{k=1}^{N}(\overline{T\mathbf{v}_k})\mathbf{v}_k\rangle| \leq c\|\mathbf{x}\|$, for any $N \in \mathbb{N}$, $\mathbf{x} \in \mathbb{V}$, where c is a constant.
(c) For $\mathbf{x} = \sum_{k=1}^{N}(\overline{T\mathbf{v}_k})\mathbf{v}_k$ in (b), show that $\sum_{k=1}^{N}|T\mathbf{v}_k|^2 \leq c^2$ for any $N \in \mathbb{N}$, and hence conclude that $\sum_{k=1}^{\infty}|T\mathbf{v}_k|^2 < \infty$.
(d) Apply (c) to show that the partial sums of $\mathbf{x}_T := \sum_{k=1}^{\infty}(\overline{T\mathbf{v}_k})\mathbf{v}_k$ is a Cauchy sequence, and hence conclude that $\mathbf{x}_T$ is in $\mathbb{V}$.

Exercise 12 Fill in the details in the proof of Theorem 1, by showing that $\|T\mathbf{x} - \sum_{|k|\leq N} c_k T\mathbf{v}_k\| \to 0$, as $N \to \infty$, and that $\sum_k \langle c_k \mathbf{v}_k, \sum_j \overline{(T\mathbf{v}_j)}\widetilde{\mathbf{v}}_j\rangle = \sum_k c_k \sum_j T\mathbf{v}_j \langle \mathbf{v}_k, \widetilde{\mathbf{v}}_j\rangle$ is valid.

Exercise 13 Let T be the functional on $\mathbb{C}^2$ or $\mathbb{R}^2$ defined by

$$T\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 2x_1 - x_2.$$

(a) Verify that T is linear.
(b) Apply each pair of dual bases, as identified in Exercise 8 on p.52 in Sect. 1.4 of Chap. 1, to formulate the representer $\mathbf{x}_T$ of T.

Exercise 14 Write down B_j^* of the following matrices B_j, and identify those that are Hermitian.

(a) $B_1 = \begin{bmatrix} 1 & i \\ 1+i & -i \end{bmatrix}$.   (b) $B_2 = \begin{bmatrix} 1 & 1-i \\ 1+i & i \end{bmatrix}$.

(c) $B_3 = \begin{bmatrix} 2 & 1+i \\ 1-i & 3 \end{bmatrix}$.   (d) $B_4 = \begin{bmatrix} -1 & 2-i \\ 2+i & -3 \end{bmatrix}$.

2.3 Eigenspaces

Corresponding to a linear transformation T from a vector space $\mathbb{V}$ to some vector space $\mathbb{W}$, the set of $\mathbf{x} \in \mathbb{V}$ which is mapped to the zero vector $\mathbf{0}$ in $\mathbb{W}$, namely:

$$\mathbb{N}_0 = \{\mathbf{x} \in \mathbb{V} : T\mathbf{x} = \mathbf{0}\} \tag{2.3.1}$$

is a subspace of $\mathbb{V}$, called the "null space" of T. Indeed, for any $\mathbf{x}, \mathbf{y} \in \mathbb{N}_0$ and scalars $a, b \in \mathbb{F}$, it follows from the linearity of T in (2.2.1) that

$$T(a\mathbf{x} + b\mathbf{y}) = aT\mathbf{x} + bT\mathbf{y} = a\,\mathbf{0} + b\,\mathbf{0} = \mathbf{0}.$$

In this section, we extend the study of the null space to "eigenspaces" for linear operators; that is, for $\mathbb{W} = \mathbb{V}$ (see Definition 2 on p.80).

Since every vector space contains the zero vector $\mathbf{0}$ (and is therefore never an "empty" set), the vector space $\mathbb{N}_0$ is said to be trivial (as opposed to being empty) if $T\mathbf{x} = \mathbf{0}$ implies $\mathbf{x} = \mathbf{0}$. In general, a vector space is said to be nontrivial if it contains at least one non-zero $\mathbf{x}$. Also, for all practical purposes, only the scalar fields $\mathbb{C}$ and $\mathbb{R}$ are used in our study of eigenspaces.

Definition 1 | **Eigenvalues** | *Let $\mathbb{V}$ be a vector space over the scalar field $\mathbb{F} = \mathbb{C}$ or $\mathbb{R}$, and let I denote the identity operator, meaning that $I\mathbf{x} = \mathbf{x}$ for all $\mathbf{x} \in \mathbb{V}$. Then a scalar $\lambda \in \mathbb{F}$ (that is, λ is a real or complex number) is called an eigenvalue of a linear operator T on $\mathbb{V}$, if the null space*

$$\mathbb{N}_\lambda = \{\mathbf{x} \in \mathbb{V} : (T - \lambda I)\,\mathbf{x} = \mathbf{0}\} \tag{2.3.2}$$

of the linear operator

$$T_\lambda = T - \lambda I$$

is nontrivial. Furthermore, if λ is an eigenvalue of T, then any non-zero $\mathbf{x} \in \mathbb{N}_\lambda$ is called an eigenvector corresponding to (or associated with) the eigenvalue λ. In addition, $\mathbb{N}_\lambda$ is called the eigenspace corresponding to λ.

| **Eigenvalues for square matrices** | In particular, if $T = A$ is an $n \times n$ square matrix, then λ is an eigenvalue of A, if there is a non-zero $\mathbf{v} \in \mathbb{C}^n$ (called a λ-eigenvector or an eigenvector associated with λ), such that $A\mathbf{v} = \lambda \mathbf{v}$; namely, the system

$$(A - \lambda I_n)\mathbf{x} = \mathbf{0} \tag{2.3.3}$$

of n linear equations has a nontrivial solution $\mathbf{x} = \mathbf{v}$. This agrees with the definition of eigenvalues in an elementary course of Linear Algebra or Matrix Theory, since the (linear) system (2.3.3) has a nontrivial solution $\mathbf{x} = \mathbf{v}$, if and only if the matrix $A - \lambda I_n$ is singular; or equivalently, $\det(\lambda I_n - A) = 0$. Let

$$P(\lambda) := \det(\lambda I_n - A). \tag{2.3.4}$$

Then $P(\lambda)$ is a polynomial of degree n, called the characteristic polynomial of the given matrix A. Thus, λ is an eigenvalue of A, if and only if λ is a (real or complex) root of its characteristic polynomial $P(\lambda)$. By the Fundamental Theorem of Algebra, if multiplicities of roots are counted, then $P(\lambda)$ has exactly n (real or complex) roots $\lambda_1, \lambda_2, \ldots, \lambda_n$; that is, A has exactly n eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$, counting multiplicities.

When a linear transformation is some square matrix, then the trace of the matrix, defined by the sum of its main diagonal entries, is given by the sum of its eigenvalues, as follows.

Definition 2 $\boxed{\textbf{Trace of matrices}}$ *The trace of a square matrix $A = \big[a_{jk}\big]_{1 \le j,k \le n}$ is defined by the sum of its main diagonal entries; namely,*

$$\mathrm{Tr}(A) = \sum_{j=1}^{n} a_{jj}.$$

Theorem 1 $\boxed{\textbf{Trace of square matrix is sum of its eigenvalues}}$ *Let A be an $n \times n$ matrix. Then*

$$\mathrm{Tr}(A) = \sum_{j=1}^{n} \lambda_j, \tag{2.3.5}$$

where λ_j, $1 \le j \le n$ (with multiplicities being listed) are the eigenvalues of A.

Let $P(\lambda)$ be the characteristic polynomial of A defined by (2.3.4). Then $P(\lambda)$ can be written as

$$P(\lambda) = \lambda^n - (a_{11} + a_{22} + \cdots + a_{nn})\lambda^{n-1} + \cdots + (-1)^n \det(A).$$

On the other hand, $P(\lambda)$ can also be written as

$$
\begin{aligned}
P(\lambda) &= (\lambda - \lambda_1)(\lambda - \lambda_2) \cdots (\lambda - \lambda_n) \\
&= \lambda^n - (\lambda_1 + \lambda_2 + \cdots + \lambda_n)\lambda^{n-1} + \cdots + (-1)^n \lambda_1 \lambda_2 \ldots \lambda_n,
\end{aligned}
$$

where $\lambda_j \in \mathbb{C}$, $1 \le j \le n$ are the eigenvalues of A. Therefore, equating the coefficient of λ^{n-1} in the above two expressions of $P(\lambda)$ yields

$$\sum_{j=1}^{n} a_{jj} = \sum_{j=1}^{n} \lambda_j. \qquad \blacksquare$$

Example 1 Determine the eigenvalues and corresponding eigenspaces of the following 2×2 matrices:

$$A_1 = \begin{bmatrix} 1 & 4 \\ 3 & 2 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 1 & -i \\ i & 2 \end{bmatrix},$$

$$A_3 = \begin{bmatrix} 1 & i \\ i & 2 \end{bmatrix}, \quad A_4 = \begin{bmatrix} 2 & 1-i \\ 1+i & 3 \end{bmatrix}.$$

Solution To compute the eigenvalues and study the eigenspaces of $A_1, \ldots, A_4$, we may compute the determinants of $A_j - \lambda I$, for $j = 1, \ldots, 4$.

(a) For the linear operator A_1 over $\mathbb{R}$, we have

$$\det(A_1 - \lambda I) = (1 - \lambda)(2 - \lambda) - 12$$
$$= \lambda^2 - 3\lambda + 2 - 12 = \lambda^2 - 3\lambda - 10,$$

so that

$$\lambda = \frac{3 \pm \sqrt{9 + 40}}{2} = \frac{3 \pm \sqrt{49}}{2} = \frac{3 \pm 7}{2},$$

which yields the eigenvalues $\lambda_1 = \dfrac{3 + 7}{2} = 5$ and $\lambda_2 = \dfrac{3 - 7}{2} = -2$ of A_1. To formulate the corresponding eigenspaces $\mathbb{N}_{\lambda_1}$ and $\mathbb{N}_{\lambda_2}$, consider the matrices

$$A_1 - \lambda_1 I = \begin{bmatrix} -4 & 4 \\ 3 & -3 \end{bmatrix} \quad \text{and} \quad A_1 - \lambda_2 I = \begin{bmatrix} 3 & 4 \\ 3 & 4 \end{bmatrix},$$

with null spaces given by

$$\mathbb{N}_{\lambda_1} = \mathrm{span}\left\{ \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right\} \quad \text{and} \quad \mathbb{N}_{\lambda_2} = \mathrm{span}\left\{ \begin{bmatrix} 4 \\ -3 \end{bmatrix} \right\},$$

respectively, since

$$[-4 \ 4] \begin{bmatrix} 1 \\ 1 \end{bmatrix} = 0 \quad \text{and} \quad [3 \ 4] \begin{bmatrix} 4 \\ -3 \end{bmatrix} = 0.$$

Observe that $\mathbb{N}_{\lambda_1}$ and $\mathbb{N}_{\lambda_2}$ are 1-dimensional vector spaces.

(b) For A_2, we have

$$\det(A_2 - \lambda I) = (1 - \lambda)(2 - \lambda) - (-i)(i)$$
$$= \lambda^2 - 3\lambda + 2 - 1 = \lambda^2 - 3\lambda + 1,$$

so that

$$\lambda = \frac{3 \pm \sqrt{9 - 4}}{2} = \frac{3 \pm \sqrt{5}}{2};$$

that is, the eigenvalues of A_2 are

$$\lambda_1 = \frac{1}{2}(3 + \sqrt{5}) \text{ and } \lambda_2 = \frac{1}{2}(3 - \sqrt{5}).$$

We leave the formulation of the eigenspaces $\mathbb{N}_{\lambda_1}$ and $\mathbb{N}_{\lambda_2}$ as an exercise (see Exercise 3).

(c) For A_3, we have

$$\det(A_3 - \lambda I) = (1 - \lambda)(2 - \lambda) - (-i)^2 = \lambda^2 - 3\lambda + 3,$$

so that

$$\lambda = \frac{3 \pm \sqrt{9 - 12}}{2} = \frac{1}{2}(3 \pm i\sqrt{3});$$

that is, the eigenvalues of A_3 are

$$\lambda_1 = \frac{1}{2}(3 + i\sqrt{3}) \text{ and } \lambda_2 = \frac{1}{2}(3 - i\sqrt{3}).$$

The formulation of the corresponding eigenspaces $\mathbb{N}_{\lambda_1}$ and $\mathbb{N}_{\lambda_2}$ is left as an exercise (see Exercise 4).

(d) For A_4, we have

$$\det(A_4 - \lambda I) = (2 - \lambda)(3 - \lambda) - (1 - i)(1 + i)$$
$$= \lambda^2 - 5\lambda + 6 - 2 = \lambda^2 - 5\lambda + 4 = (\lambda - 4)(\lambda - 1),$$

so that the eigenvalues of A_4 are $\lambda_1 = 1$ and $\lambda_2 = 4$. Observe that

$$A_4 - \lambda_1 I = \begin{bmatrix} 2 - \lambda_1 & 1 - i \\ 1 + i & 3 - \lambda_1 \end{bmatrix} = \begin{bmatrix} 1 & 1 - i \\ 1 + i & 2 \end{bmatrix}$$
$$= \begin{bmatrix} 1 & 1 - i \\ 1 + i & (1 + i)(1 - i) \end{bmatrix}$$

and

$$A_4 - \lambda_2 I = \begin{bmatrix} 2 - \lambda_2 & 1 - i \\ 1 + i & 3 - \lambda_2 \end{bmatrix} = \begin{bmatrix} -2 & 1 - i \\ 1 + i & -1 \end{bmatrix}$$
$$= \begin{bmatrix} (i - 1)(1 + i) & (i - 1)(-1) \\ 1 + i & -1 \end{bmatrix};$$

so that the two eigenspaces are given by

$$\mathbb{N}_{\lambda_1} = \text{span}\left\{ \begin{bmatrix} 1 - i \\ -1 \end{bmatrix} \right\} \text{ and } \mathbb{N}_{\lambda_2} = \text{span}\left\{ \begin{bmatrix} 1 \\ 1 + i \end{bmatrix} \right\},$$

since

$$[1 \quad 1 - i]\begin{bmatrix} 1 - i \\ -1 \end{bmatrix} = 0 \text{ and } [1 + i \quad -1]\begin{bmatrix} 1 \\ 1 + i \end{bmatrix} = 0.$$

∎

Let us now turn to the study of linear operators that are not necessarily square matrices.

Example 2 Let $\mathbb{V} = C^\infty(-\infty, \infty)$ be the vector space of infinitely differentiable real-valued functions on $\mathbb{R} = (-\infty, \infty)$ over $\mathbb{R}$; that is, $f \in \mathbb{V}$ if and only if f is real-valued and the derivatives $f, f', f'', \ldots$ are continuous functions on $\mathbb{R}$. Consider the linear operator T defined by taking the second derivative. More precisely, let

$$Tf = -D^2 f = -f''.$$

Determine the eigenvalues of T and exhibit their corresponding eigenspaces.

Solution Observe that since $\mathbb{V} = C^\infty(-\infty, \infty)$ is not a finite-dimensional space, T cannot be represented as a (finite) square matrix. On the other hand, it is easy to verify that T is a linear operator on $\mathbb{V}$, and the eigenspace corresponding to λ is the null space of $T - \lambda I$; namely, the vector space of all $\mathbf{x} = f(t)$, for which

$$T\mathbf{x} - \lambda\mathbf{x} = -f''(t) - \lambda f(t) = \mathbf{0}$$

(see Definition 1). In other words, we have to solve the ordinary differential equation

$$y'' + \lambda y = 0 \tag{2.3.6}$$

(for nontrivial solutions) to compute the eigenvalues λ of T. For convenience, the discussion below is divided into three separate cases.

  (i) Consider $\lambda = 0$. Then the solution of (2.3.6) is simply

$$y = a + bt, \quad a, b \in \mathbb{R}.$$

Hence, $\lambda = 0$ is an eigenvalue of $T = -D^2$ and the corresponding eigenspace is the null space of T, namely:

$$\mathbb{N}_0 = \text{span}\{1, t\}, \tag{2.3.7}$$

the space of linear polynomials with algebraic basis $\{1, t\}$.

  (ii) Let $\lambda < 0$, and write $\lambda = -\mu^2$, where $\mu > 0$. Then the differential equation (2.3.6) has two linearly independent solutions:

$$f_1(t) = e^{-\mu t}, \quad f_2(t) = e^{\mu t}. \tag{2.3.8}$$

Hence, every negative number $\lambda = -\mu^2$ is an eigenvalue of $-D^2$, with corresponding eigenspace

$$\mathbb{N}_\lambda = N_{-\mu^2} = \text{span}\{e^{-\mu t}, e^{\mu t}\}. \tag{2.3.9}$$

(iii) Let $\lambda > 0$, and write $\lambda = \mu^2$, where $\mu > 0$. Then the differential equation (2.3.6) has two linearly independent solutions

$$g_1(t) = \cos \mu t, \quad g_2(t) = \sin \mu t. \tag{2.3.10}$$

Hence, every positive number $\lambda = \mu^2$ is an eigenvalue of $T = -D^2$, with corresponding eigenspace

$$\mathbb{N}_\lambda = \mathbb{N}_{\mu^2} = \text{span}\{\cos \mu t, \sin \mu t\}. \tag{2.3.11}$$

∎

Remark 1 $\boxed{\textbf{Eigenfunction}}$ For function spaces, eigenvectors are also called eigenfunctions. For example, for a given linear (differential or integral) operator T, a nonzero function f that satisfies $Tf = \lambda f$, where λ is a scalar, is called a λ-eigenfunction or an eigenfunction associated with the eigenvalue λ of T.   ∎

In the following example, we replace the infinite interval $(-\infty, \infty)$ in Example 2 by a finite interval and impose certain boundary conditions.

Example 3   Let $J = [0, c]$, where $c > 0$, and consider the set $C_0^\infty(J)$ of infinitely differentiable functions $f(t)$ on J with the two-point boundary condition $f'(0) = f'(c) = 0$. Determine the eigenvalues and corresponding eigenspaces of the linear operator $T = -D^2$ on $C_0^\infty(J)$, as introduced in Example 2.

Solution   It is clear that $C_0^\infty(J)$ is a vector space, since for all $a, b \in \mathbb{R}$ and $f, g \in C_0^\infty(J)$, the function $h = af + bg$ is in $C^\infty(J)$ and satisfies the two-point boundary condition

$$h'(0) = af'(0) + bg'(0) = 0,$$
$$h'(c) = af'(c) + bg'(c) = 0.$$

We again consider three cases.

(i) Consider $\lambda = 0$. Then the solution of the two-point boundary value problem

$$\begin{cases} y'' \quad = 0; \\ y'(0) = y'(c) = 0 \end{cases}$$

has the general solution $y = q, q \in \mathbb{R}$. Hence, $\lambda_0 = 0$ is an eigenvalue of $T = -D^2$ and the corresponding eigenspace is

$$N_{\lambda_0} = N_0 = \text{span}\{1\}. \tag{2.3.12}$$

(ii) Consider $\lambda = -\mu^2$, $\mu > 0$. Then the solution of the two-point boundary value problem

$$\begin{cases} y'' - \mu^2 y & = 0; \\ y'(0) = y'(c) = 0 \end{cases}$$

is given by $y = f(t) = ae^{-\mu t} + be^{\mu t}$ μ with $f'(0) = (-a + b)\mu = 0$, or $a = b$; and in addition, $f'(c) = (-ae^{-\mu c} + be^{\mu c}) = 0$, or $a = be^{2\mu c}$. Hence, $a = b = 0$. That is, T has no negative eigenvalue (since $N_\lambda = N_{-\mu^2} = \{0\}$).

(iii) Consider $\lambda = \mu^2$, $\mu > 0$. Then the solution of the two-point boundary value problem

$$\begin{cases} y'' + \mu^2 y & = 0; \\ y'(0) = y'(c) = 0 \end{cases}$$

is given by $y = f(t) = a \cos \mu t + b \sin \mu t$, with

$$f'(0) = -a\mu \sin 0 + b\mu \cos 0 = b\mu = 0, \ \text{ or } b = 0;$$

and in addition,

$$f'(c) = -a\mu \sin \mu c = 0, \ \text{ or } \mu c = k\pi,$$

where k is any integer. Therefore, the eigenvalues of $T = -D^2$ are

$$\lambda_k = \left(\frac{k\pi}{c}\right)^2, \ k = 0, 1, 2, \ldots, \tag{2.3.13}$$

since $\lambda_0 = 0$ has been taken care of in Case (i). The eigenspace N_{λ_k} corresponding to each λ_k, $k = 0, 1, 2, \ldots$, is given by

$$N_{\lambda_k} = \text{span}\left\{\cos \frac{\pi k t}{c}\right\}, \tag{2.3.14}$$

which includes (2.3.12), because $\cos 0 = 1$.

∎

Remark 2 In Sect. 6.5 of Chap. 6, the result in Example 3 will be applied as "spatial" basis functions to solving the heat diffusion partial differential equations, after the spatial and time components are separated. ∎

We now return to Example 1 and observe that the eigenvalues of the Hermitian (self-adjoint) matrices A_2 and A_4 (that is, $A_2^* = \overline{A_2}^T = A_2$ and $A_4^* = \overline{A_4}^T = A_4$) are real, but the eigenvalues of the symmetric matrix A_3 (with $A_3^T = A_3$, but $A_3^* \neq A_3$) are not real numbers. This example serves as a motivation of the following result.

Theorem 2 | **Eigenvalues of self-adjoint operators are real** | *Eigenvalues of a self-adjoint (or Hermitian) linear operator T on any inner-product space over $\mathbb{F} = \mathbb{C}$ or $\mathbb{R}$ are real numbers.*

Proof The proof is quite simple. Indeed, $T^* = T$ implies that

$$
\begin{aligned}
\lambda \langle \mathbf{x}_\lambda, \mathbf{x}_\lambda \rangle &= \langle \lambda \mathbf{x}_\lambda, \mathbf{x}_\lambda \rangle = \langle T \mathbf{x}_\lambda, \mathbf{x}_\lambda \rangle \\
&= \langle \mathbf{x}_\lambda, T^* \mathbf{x}_\lambda \rangle = \langle \mathbf{x}_\lambda, T \mathbf{x}_\lambda \rangle \\
&= \langle \mathbf{x}_\lambda, \lambda \mathbf{x}_\lambda \rangle = \overline{\lambda} \langle \mathbf{x}_\lambda, \mathbf{x}_\lambda \rangle,
\end{aligned}
$$

where λ is any eigenvalue and $\mathbf{x}_\lambda \in \mathbb{N}_\lambda$ is a corresponding eigenvector. Now, being an eigenvector, $\mathbf{x}_\lambda \neq \mathbf{0}$ or $\langle \mathbf{x}_\lambda, \mathbf{x}_\lambda \rangle = \|\mathbf{x}_\lambda\|^2 > 0$, so that $\lambda = \overline{\lambda}$. Hence, λ is real. ∎

Theorem 3 | **Orthogonality of eigenspaces** | *Let T be a self-adjoint (or Hermitian) operator on an inner-product space $\mathbb{V}$ over $\mathbb{F} = \mathbb{C}$ or $\mathbb{R}$. Suppose that λ_1 and λ_2 are two distinct eigenvalues of T. Then the two corresponding eigenspaces $\mathbb{N}_{\lambda_1}$ and $\mathbb{N}_{\lambda_2}$ are orthogonal to each other; that is, $\langle \mathbf{x}, \mathbf{y} \rangle = 0$ for all $\mathbf{x} \in \mathbb{N}_{\lambda_1}$ and $\mathbf{y} \in \mathbb{N}_{\lambda_2}$.*

Proof The proof is similar to that of Theorem 2 . Let $\mathbf{x} \in \mathbb{N}_{\lambda_1}$ and $\mathbf{y} \in \mathbb{N}_{\lambda_2}$. Then

$$
\begin{aligned}
\lambda_1 \langle \mathbf{x}, \mathbf{y} \rangle &= \langle \lambda_1 \mathbf{x}, \mathbf{y} \rangle = \langle T \mathbf{x}, \mathbf{y} \rangle \\
&= \langle \mathbf{x}, T^* \mathbf{y} \rangle = \langle \mathbf{x}, T \mathbf{y} \rangle = \langle \mathbf{x}, \lambda_2 \mathbf{y} \rangle \\
&= \overline{\lambda_2} \langle \mathbf{x}, \mathbf{y} \rangle = \lambda_2 \langle \mathbf{x}, \mathbf{y} \rangle,
\end{aligned}
$$

since λ_2 is real according to Theorem 2. Therefore, we have

$$
(\lambda_1 - \lambda_2) \langle \mathbf{x}, \mathbf{y} \rangle = 0,
$$

which implies that $\langle \mathbf{x}, \mathbf{y} \rangle = 0$, since $\lambda_1 \neq \lambda_2$.   ∎

Remark 3 | **Orthogonality of left and right eigenvectors** | If $T = A$ is a Hermitian matrix, then by Theorems 2 and 3, all eigenvalues of A are real, and two (right) eigenvectors associated with two distinct eigenvalues are orthogonal. Corresponding to the second property, we have the following result for general real matrices.

For any eigenvalue-eigenvector pair $(\lambda, \mathbf{v})$ of a square matrix A, the row vector $\mathbf{u} = \mathbf{v}^T$ satisfies

$$
\mathbf{u} A^T = \lambda \mathbf{u}.
$$

We will call $\mathbf{u}$ a left eigenvector of the matrix A^T. For convenience, we may also call the eigenvector $\mathbf{v}$ a right eigenvector of A. If A is a real matrix, then since A^T is the adjoint of A, the above proof of Theorem 3 can be adopted to show that a left eigenvector corresponding to some eigenvalue λ_1 and a right eigenvector corresponding

to an eigenvalue λ_2 are orthogonal to each other, provided that $\lambda_1 \neq \bar{\lambda}_2$. This result extends Theorem 3 to non-self-adjoint (but real) matrices (see Exercise 17). ∎

Example 4 Let A_1 and A_4 be the matrices

$$A_1 = \begin{bmatrix} 1 & 4 \\ 3 & 2 \end{bmatrix}, \quad A_4 = \begin{bmatrix} 2 & 1-i \\ 1+i & 3 \end{bmatrix}$$

considered in Example 1. Apply A_1 to illustrate the above Remark 3, and A_4 to illustrate Theorem 3.

Solution For A_1, recall from Example 1 that the eigenvectors corresponding to the eigenvalues $\lambda_1 = 5$ and $\lambda_2 = -2$ are

$$\mathbf{v}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \mathbf{v}_2 = \begin{bmatrix} 4 \\ -3 \end{bmatrix},$$

respectively. Since A_1 is not self-adjoint, we do not expect $\mathbf{v}_1$ to be orthogonal to $\mathbf{v}_2$, although the two eigenvalues are different. However, for the eigenvalue $\lambda_1 = 5$, we have

$$A_1 - \lambda_1 I_2 = \begin{bmatrix} -4 & 4 \\ 3 & -3 \end{bmatrix};$$

and it is easy to verify that the row vector $\mathbf{u} = [3, 4]$ is a left eigenvector associated with $\lambda_1 = 5$. From Remark 3, since $\lambda_1 \neq \bar{\lambda}_2$, this vector $\mathbf{u}$ must be orthogonal to the (right) eigenvector $\mathbf{v}_2$ associated with the eigenvalue $\lambda_2 = -2$. This is indeed the case, namely:

$$\mathbf{u}\bar{\mathbf{v}}_2 = [3, 4] \begin{bmatrix} 4 \\ -3 \end{bmatrix} = 12 - 12 = 0.$$

As to the matrix A_4, it is clear that it is self-adjoint. From Example 1, we already know that the eigenvectors are (non-zero) constant multiples of

$$\mathbf{v}_1 = \begin{bmatrix} 1-i \\ -1 \end{bmatrix}, \mathbf{v}_2 = \begin{bmatrix} 1 \\ 1+i \end{bmatrix}$$

associated with the corresponding eigenvalues $\lambda_1 = 1$ and $\lambda_2 = 4$. Since $\lambda_1 \neq \lambda_2$ and A_4 is Hermitian, we expect $\mathbf{v}_1$ and $\mathbf{v}_2$ to be orthogonal to each other, as assured by Theorem 3. That this is indeed the case can be verified by direct calculation, namely:

$$\langle \mathbf{v}_1, \mathbf{v}_2 \rangle = \mathbf{v}_1^T \bar{\mathbf{v}}_2 = [1-i, -1] \begin{bmatrix} 1 \\ 1-i \end{bmatrix} = 1 - i - (1-i) = 0.$$

∎

Example 5 As a continuation of Example 3, where the eigenvalues of the differential linear operator $T = -D^2$ are

$$\lambda_k = \left(\frac{k\pi}{c}\right)^2, \ k = 0, 1, 2, \ldots,$$

and their corresponding eigenspaces given by

$$\mathbb{N}_{\lambda_k} = \text{span}\left\{\cos\frac{\pi kt}{c}\right\}, \ k = 0, 1, 2, 3, \ldots,$$

in view of (2.3.12) and (2.3.14), verify that T is self-adjoint on $C_0^\infty(J)$ and that the eigenspaces $\mathbb{N}_{\lambda_k}$ are mutually orthogonal.

Solution For $f, g \in C_0^\infty(J)$, it follows by the two-point boundary condition and integration by parts that

$$
\begin{aligned}
\langle Tf, g \rangle &= \int_0^c \left(-f''(t)\right)\overline{g(t)}dt \\
&= -\left[f'(t)\overline{g(t)}\right]_0^c + \int_0^c f'(t)\overline{g'(t)}dt \\
&= -\left[f'(t)\overline{g(t)}\right]_0^c + \left[f(t)\overline{g'(t)}\right]_0^c - \int_0^c f(t)\overline{g''(t)}dt \\
&= \int_0^c f(t)\left(-\overline{g''(t)}\right)dt = \langle f, Tg \rangle,
\end{aligned}
$$

since $f'(0) = f'(c) = 0$ and $g'(0) = g'(c) = 0$. That is, T is self-adjoint.

In addition, observe that for all $j, k = 0, 1, \ldots$, we have

$$
\begin{aligned}
\langle \cos\frac{\pi jt}{c}, \cos\frac{\pi kt}{c}\rangle &= \int_0^c \cos\frac{\pi jt}{c}\cos\frac{\pi kt}{c}dt \\
&= \frac{c}{\pi}\int_0^\pi \cos jt \cos kt\, dt \\
&= \frac{c}{2\pi}\left\{\int_0^\pi \cos(j-k)t\, dt + \int_0^\pi \cos(j+k)t\, dt\right\} \\
&= \frac{c}{2\pi}\left\{\left[\frac{\sin(j-k)t}{j-k}\right]_0^\pi + \left[\frac{\sin(j+k)t}{j+k}\right]_0^\pi\right\} = 0,
\end{aligned}
$$

which is valid whenever $j \neq k$.   ∎

Of course an equivalent statement of Theorem 3 can be formulated as follows: "For self-adjoint operators on an inner-product space, eigenvectors corresponding to different eigenvalues are orthogonal to one another". Hence, to derive an orthonormal family from all the eigenvectors, it is sufficient to replace the algebraic basis of every

eigenspace by an orthonormal basis. The procedure to achieve this goal is to apply the Gram-Schmidt orthogonalization process introduced in Sect. 1.4 of Chap. 1.

Next, we observe that the self-adjoint matrices A_2 and A_4 in Example 1 also satisfy the property:

$$\langle \mathbf{x}, A\mathbf{x} \rangle \geq 0, \ \text{for all } \mathbf{x} \in \mathbb{C}^2, \tag{2.3.15}$$

where $A = A_2, A_4$. We only verify the case $A = A_2$ and leave the case $A = A_4$ as an exercise (see Exercise 9). Indeed, for $\mathbf{x} = [a, b]$, where $a, b \in \mathbb{C}$, we have

$$\langle \mathbf{x}, A_2\mathbf{x} \rangle = [a \ b] \overline{\begin{bmatrix} 1 & -i \\ i & 2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix}} = [a \ b] \overline{\begin{bmatrix} a - ib \\ ia + 2b \end{bmatrix}}$$

$$= a(\bar{a} + i\bar{b}) + b(-i\bar{a} + 2\bar{b})$$
$$= a\bar{a} + ia\bar{b} - i\bar{a}b + 2b\bar{b}$$
$$= (a - ib)\overline{(a - ib)} + |b|^2$$
$$= |a - ib|^2 + |b|^2 \geq 0.$$

Motivated by this example, we introduce the concept of

Definition 3 | **Self-adjoint positive semi-definite (spd) operators** | *A self-adjoint linear operator T on an inner-product space $\mathbb{V}$ over the scalar field $\mathbb{F} = \mathbb{C}$ or $\mathbb{R}$ is said to be positive semi-definite if $\langle T\mathbf{x}, \mathbf{x} \rangle \geq 0$ for all $\mathbf{x} \in \mathbb{V}$. Furthermore, an spd operator is called positive definite, if $\langle T\mathbf{x}, \mathbf{x} \rangle = 0$ implies that $\mathbf{x} = 0$.*

In the literature, "spd" stands for both "self-adjoint positive semi-definite" and "self-adjoint positive definite" operators.

Example 6 As discussed above, the matrices A_2 and A_4 in Example 1 are spd operators on the inner-product space $\mathbb{C}^2$ over the scalar field $\mathbb{C}$.

Example 7 The linear operator $T = -D^2$ in Example 3 is spd on the inner-product space $C_0^\infty(J)$ over $\mathbb{R}$, where $J = [0, c]$.

Solution Indeed, for any $f \in C_0^\infty(J)$, an application of integration by parts, as in Example 5 with $g = f$, immediately yields:

$$\langle Tf, f \rangle = -\int_0^c f''(t)\overline{f(t)}dt$$
$$= -\left[f'(t)\overline{f(t)}\right]_0^c + \int_0^c f'(t)\overline{f'(t)}dt$$
$$= \int_0^c |f'(t)|^2 dt \geq 0.$$

∎

There is a distinction between the above two examples. While $\langle A_2 \mathbf{x}, \mathbf{x} \rangle = |a - ib|^2 + |b|^2 = 0$ implies $a = b = 0$ in Example 6, the differential operator $T = -D^2$ in Example 7 annihilates the nonzero constant function q, in that $\langle Tq, q \rangle = \int_0^c \left| \dfrac{d}{dt} q \right|^2 dt = 0$ although $q \neq 0$. Hence, A_2 is positive definite while $T = -D^2$ is positive semi-definite, though both operators are called spd, being self-adjoint operators.

Theorem 4 $\boxed{\textbf{Eigenvalues of spd operators are } \geq 0}$ *Let T be an spd linear operator on an inner-product space over $\mathbb{F} = \mathbb{C}$ or $\mathbb{R}$. Then all eigenvalues λ of T are non-negative.*

The proof is evident. Indeed, if λ is an eigenvalue of an spd operator T with corresponding eigenvector $\mathbf{x}_\lambda$, then

$$\lambda \langle \mathbf{x}_\lambda, \mathbf{x}_\lambda \rangle = \langle T\mathbf{x}_\lambda, \mathbf{x}_\lambda \rangle \geq 0$$

so that $\lambda \geq 0$, since $\langle \mathbf{x}_\lambda, \mathbf{x}_\lambda \rangle = \|\mathbf{x}_\lambda\|^2 > 0$.   ∎

However, computation of eigenvalues is not an easy task in general. Even for square matrices A, computing the determinant of $A - \lambda I$, followed by finding the roots (also called zeros) of the characteristic polynomial $\det(A - \lambda I)$, is a very difficult task for square matrices of high dimensions. The task would not seem to be feasible for linear operators on infinite-dimensional inner-product spaces in general. Fortunately, there are numerical algorithms available in the literature, particularly for spd linear operators T, for which $0 \leq \lambda \langle \mathbf{x}, \mathbf{x} \rangle = \langle \lambda \mathbf{x}, \mathbf{x} \rangle = \langle T\mathbf{x}, \mathbf{x} \rangle$ (provided that $\lambda \geq 0$ is an eigenvalue and $\mathbf{x}$ a corresponding eigenvector), so that

$$\lambda = \frac{\langle T\mathbf{x}, \mathbf{x} \rangle}{\|\mathbf{x}\|^2}.$$

From the information that all eigenvalues of an spd linear operator T are non-negative, it is at least intuitively clear that the smallest eigenvalue is given by

$$\lambda_0 = \inf_{\mathbf{x} \neq 0} \frac{\langle T\mathbf{x}, \mathbf{x} \rangle}{\|\mathbf{x}\|^2}, \tag{2.3.16}$$

where "inf" stands for "infimum", which means the "greatest lower bound". The quotient $\dfrac{\langle T\mathbf{x}, \mathbf{x} \rangle}{\|\mathbf{x}\|^2}$ in (2.3.16) is called the "Rayleigh quotient" for the spd operator T.

For a subspace $\mathbb{W} \subset \mathbb{V}$, let $\mathbb{W}^\perp$ denote the orthogonal complement of $\mathbb{W}$ in $\mathbb{V}$, as defined by (1.3.11) on p.28. Hence, it follows from Theorem 3 that the second smallest eigenvalue $\widetilde{\lambda}_1$ satisfies

$$\widetilde{\lambda}_1 = \inf \left\{ \frac{\langle T\mathbf{x}, \mathbf{x} \rangle}{\|\mathbf{x}\|^2} : \mathbf{0} \neq \mathbf{x} \in \mathbb{N}_{\lambda_0}^\perp \right\}. \tag{2.3.17}$$

Observe that since the infimum in (2.3.17) is taken over a proper subset of $\mathbb{V}$ as compared to the set $\mathbb{V}$ in (2.3.16), we have $\widetilde{\lambda}_1 > \lambda_0$. In general, by listing only the distinct eigenvalues of the spd linear operator T in increasing order; namely,

$$0 \leq \widetilde{\lambda}_0 < \widetilde{\lambda}_1 < \widetilde{\lambda}_2 < \ldots$$

(where $\widetilde{\lambda}_0 := \lambda_0$), we have

$$\widetilde{\lambda}_{n+1} = \inf\left\{\frac{\langle T\mathbf{x}, \mathbf{x}\rangle}{\|\mathbf{x}\|^2} : \mathbf{0} \neq \mathbf{x} \in \mathbb{N}_{\widetilde{\lambda}_j}^\perp, j = 0, \ldots, n\right\}. \qquad (2.3.18)$$

The minimization model in (2.3.16)–(2.3.18) provides a platform for computing numerical approximations of the distinct eigenvalues of spd operators. In the above discussion, we use the notation $\widetilde{\lambda}_j$ instead of λ_j to emphasize that $\widetilde{\lambda}_0, \widetilde{\lambda}_1, \ldots$ are distinct.

Remark 4 Observe that the minimization model (2.3.16)–(2.3.18) allows for the computation of infinitely many eigenvalues for a general spd operator in increasing order, starting with $\lambda_0 \geq 0$. However, if the spd operator is an $n \times n$ matrix A_n, then we may modify (2.3.16)–(2.3.18) to allow for computation of the eigenvalues of A_n in decreasing order, namely:

$$\lambda_1 = \sup\left\{\frac{\langle A_n\mathbf{x}, \mathbf{x}\rangle}{\|\mathbf{x}\|^2} : \mathbf{x} \neq \mathbf{0}\right\} \qquad (2.3.19)$$

(see Theorem 1 and Theorem 2 to be derived in Sect. 3.2 of the next chapter). Furthermore, in order to match the indices with those of the eigenvectors, multiplicities of the eigenvalues must be taken into consideration. To do so, we change the above indexing notation to

$$\lambda_1 = \cdots = \lambda_{r_1} > \lambda_{r_1+1} = \cdots = \lambda_{r_2+r_1} > \cdots > \cdots = \lambda_n \geq 0,$$

where $r_1 = \dim \mathbb{N}_{\lambda_1}, r_2 = \dim \mathbb{N}_{\lambda_{r_1+1}}, \ldots$, and

$$\lambda_{r_1+1} = \sup\left\{\frac{\langle A_n\mathbf{x}, \mathbf{x}\rangle}{\|\mathbf{x}\|^2} : \mathbf{0} \neq \mathbf{x} \in \mathbb{N}_{\lambda_1}^\perp\right\};$$

$$\lambda_{r_1+r_2+1} = \sup\left\{\frac{\langle A_n\mathbf{x}, \mathbf{x}\rangle}{\|\mathbf{x}\|^2} : \mathbf{0} \neq \mathbf{x} \in \mathbb{N}_{\lambda_{r_1+1}}^\perp \cup \mathbb{N}_{\lambda_{r_1+1}}^\perp\right\}$$

etc., and finally

$$\lambda_n = \sup\left\{\frac{\langle A_n\mathbf{x}, \mathbf{x}\rangle}{\|\mathbf{x}\|^2} : \mathbf{0} \neq \mathbf{x} \in \mathbb{N}_\lambda^\perp, \ \lambda > \lambda_n\right\} = \inf_{\mathbf{x} \neq 0} \frac{\langle A_n\mathbf{x}, \mathbf{x}\rangle}{\|\mathbf{x}\|^2}, \qquad (2.3.20)$$

where (2.3.16) is applied to formulate the last quantity in (2.3.20). ∎

Remark 5 Let A_n be an $n \times n$ spd matrix, with eigenvalues $\lambda_1, \ldots, \lambda_n$, listed in non-increasing order according to multiplicities as determined by the dimensions of the eigenspaces (see Remark 4), so that

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0.$$

According to Theorem 3 and by applying the Gram-Schmidt orthonormalization process studied in Sect. 1.4 of Chap. 1, there is an orthonormal basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ of $\mathbb{C}^n$, such that

$$(\lambda_1, \mathbf{v}_1), (\lambda_2, \mathbf{v}_2), \ldots, (\lambda_n, \mathbf{v}_n)$$

are eigenvalue-eigenvector pairs. Hence,

$$A_n \mathbf{v}_1 = \lambda_1 \mathbf{v}_1, \, A_n \mathbf{v}_2 = \lambda_2 \mathbf{v}_2, \ldots, A_n \mathbf{v}_n = \lambda_n \mathbf{v}_n,$$

and in (compact) matrix formulation,

$$A_n V = V \begin{bmatrix} \lambda_1 & & O \\ & \ddots & \\ O & & \lambda_n \end{bmatrix}, \tag{2.3.21}$$

where V denotes the $n \times n$ matrix $[\mathbf{v}_1 \ldots \mathbf{v}_n]$ with the jth column being the vector $\mathbf{v}_j$.   ∎

Remark 6 In the next section, the formulation (2.3.21) is extended from spd matrices to "normal" matrices and provides the so-called "spectral decomposition" of all normal matrices.   ∎

Remark 7 If X is an $m \times n$ rectangular matrix, then $A_n = \overline{X}^T X$ is an $n \times n$ spd matrix (see Exercise 15). Hence, we may apply Theorem 3 to each eigenspace of A_n to obtain the formulation (2.3.21). Since $\lambda_j \geq 0$ for $j = 1, \ldots, n$, we may also write $\lambda_j = \sigma_j^2$, where $\sigma_j \geq 0$. Then $\sigma_1 \geq \ldots \geq \sigma_n \geq 0$ are called the "singular values" of the $m \times n$ matrix X. Note that if $n > m$, then it is necessary that, $\sigma_{m+1} = \ldots = \sigma_n = 0$. In Sect. 3.1 of Chap. 3, we will apply this result to introduce and study the notion of singular value decomposition of the matrix X.   ∎

Exercises

Exercise 1 Determine the null spaces of the following linear transformations.

(a) $T_1 = \begin{bmatrix} 1 & -1 & 1 \\ 2 & 1 & 0 \\ 1 & 2 & -1 \end{bmatrix} : \mathbb{R}^3 \to \mathbb{R}^3$.

(b) $T_2 = \begin{bmatrix} i & 1+i & 3 & -1+3i \\ 2i & 2i & 5 & 2i \\ 0 & 2 & 1 & -2+4i \end{bmatrix} : \mathbb{C}^3 \to \mathbb{C}^3$.

(c) $(T_3 f)(x) = f''(x) - 2f'(x)$: $C^2[0, 1] \to C[0, 1]$, where $C^2[0, 1]$ denotes the vector space of functions f defined on $[0, 1]$ such that f, f', f'' are continuous on $[0, 1]$.

(d) $(T_4 f)(x) = f''(x) + f(x)$: $C^2[0, 1] \to C[0, 1]$, where $C^2[0, 1]$ is defined as in (c).

Exercise 2 Compute the eigenvalues of the following operators:

(a) $C_1 = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}$,

(b) $C_2 = \begin{bmatrix} 0 & 2 \\ -1 & 1 \end{bmatrix}$,

(c) $C_3 = \begin{bmatrix} 1 & i \\ -i & 3 \end{bmatrix}$,

(d) $C_4 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & -3 \end{bmatrix}$.

Exercise 3 Determine the two eigenspaces of the matrix A_2 in Example 1.

Exercise 4 Determine the two eigenspaces of the matrix A_3 in Example 1.

Exercise 5 Determine the eigenspaces of $C_1, \ldots, C_4$ in Exercise 2 by exhibiting their bases.

Exercise 6 Let $\mathbb{V} = C^\infty(-\infty, \infty)$ be a subspace of $\tilde{L}_2(-\infty, \infty)$, as defined in Example 2. Determine the eigenvalues of the following linear operators on $\mathbb{V}$:

(a) $T_1 f = 3f'$,
(b) $T_2 f = -(f'' + f)$,
(c) $T_3 f = -f'' + 2f$,
(d) $T_4 f = f' - 4f$.

Exercise 7 Let $C_0^\infty(J)$, $J = [0, c]$ be the subspace of $\tilde{L}_2(J)$ as defined in Example 3. Determine the eigenvalues, if any, of the linear operators $T_1, \ldots, T_4$, in Exercise 6, on $C_0^\infty(J)$. For those that have eigenvalues, determine the corresponding eigenspaces.

Exercise 8 As a continuation of Exercise 6, determine all the eigenvalues of the operator $Tf = f^{(4)}$ on the subspace $C^\infty(-\infty, \infty)$ of the inner product space $\tilde{L}_2(-\infty, \infty)$. In particular, exhibit a basis of the null space of T (that is, the eigenspace of the zero eigenvalues).

Exercise 9 Verify that the matrix $A = A_4$ in Example 1 satisfies (2.3.15).

Exercise 10 Consider the subspace

$$\mathbb{W}_0 = \{f \in C^\infty[-1, 1] : f''(-1) = f'''(-1) = f''(1) = f'''(1) = 0\}$$

of the inner-product space $\widetilde{L}_2[-1, 1]$, and the operator T on $\mathbb{W}_0$ defined by $Tf = f^{(4)}$ for $f \in \mathbb{W}_0$. Show that T is a self-adjoint operator.

Exercise 11 Let $\lambda = \nu^4$, where $\nu > 0$, and set $f_1(x) = \cos \nu x$, $f_2(x) = \sin \nu x$, $f_3(x) = \cosh \nu x$, and $f_4(x) = \sinh \nu x$.

(a) Show that f_1, f_2, f_3, f_4 are linearly independent.
(b) Show that f_1, f_2, f_3, f_4 are eigenvectors of the operator T defined by $Tf = f^{(4)}$ associated with the eigenvalue $\lambda = \nu^4$.
(c) Let $g_\lambda(x) = c_1 f_1(x) + c_2 f_2(x) + c_3 f_3(x) + c_4 f_4(x)$. Verify that for each $\nu = k\pi$, where k is an arbitrary integer $\neq 0$, (λ, g_λ) is an eigenvalue-eigenvector pair of the linear operator T on the inner-product subspace $\mathbb{W}_0$ in Exercise 10, if and only if $c_3 = c_4 = 0$.
(d) Repeat the problem in (c) for each $\nu = (k + \frac{1}{2})\pi$, where k is an arbitrary integer.

Exercise 12 As a continuation of (c) and (d) in Exercise 11, determine the eigenspaces of T for the eigenvalues $\lambda = (k\pi)^4$, where k is an arbitrary non-zero integer.

Exercise 13 Repeat Exercise 12 for the eigenvalues $\lambda = \left((k + \frac{1}{2})\pi\right)^4$, where k is an arbitrary integer.

Exercise 14 Verify (2.3.16) for the matrix $A = A_4$ in Example 1. Hence, A_4 is an spd (self-adjoint positive definite) linear operator (see Definition 3).

Exercise 15 Let X be any $m \times n$ (rectangular) matrix of complex numbers, and consider the $n \times n$ and $m \times m$ square matrices

$$A_n = \overline{X}^T X, \ B_m = X\overline{X}^T.$$

Show that both A_n and B_m are spd (self-adjoint positive semi-definite) matrices.

Exercise 16 Compute the eigenvalues and determine the corresponding eigenspaces of the matrix

$$A = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Then apply the Gram-Schmidt process studied in Sect. 1.4 of Chap. 1 to compute an orthonormal basis of $\mathbb{R}^3$ in terms of the eigenvectors of A.

Exercise 17 Let A be an $n \times n$ real matrix, and $\mathbf{u} \in \mathbb{C}^{1,n}$, $\mathbf{v} \in \mathbb{C}^{n,1}$ be vectors that satisfy $A\mathbf{v} = \lambda_1 \mathbf{v}$, $\mathbf{u}A = \lambda_2 \mathbf{u}$. Prove that if $\lambda_1 \neq \overline{\lambda}_2$, then $\mathbf{u}$ and $\mathbf{v}$ are orthogonal to each other; that is, $\langle \mathbf{u}, \mathbf{v} \rangle = \sum_{k=1}^n u_k \overline{v_k} = 0$.

Exercise 18 Show that for any $n \times n$ matrices A and B, $\text{Tr}(AB) = \text{Tr}(BA)$.

2.4 Spectral Decomposition

Loosely speaking, the set of all eigenvalues of a linear operator T on an inner-product space $\mathbb{V}$ is called the spectrum $\sigma(T)$ of T. Indeed, this is a correct statement if $\mathbb{V}$ is finite-dimensional, but $\sigma(T)$ could be a larger set in general.

Definition 1 $\boxed{\textbf{Spectrum}}$ *Let T be a linear operator on an inner-product space $\mathbb{V}$ over $\mathbb{F} = \mathbb{C}$ or $\mathbb{R}$. Then the spectrum of T, denoted by $\sigma(T)$, is the set of $\lambda \in \mathbb{F}$, for which the operator*

$$T_\lambda := T - \lambda I \tag{2.4.1}$$

is singular (*that is, not invertible*).

Example 1 Recall from the previous section that $\lambda \in \mathbb{C}$ is an eigenvalue of a matrix $A \in \mathbb{C}^{n,n}$, if it is a root of the characteristic polynomial

$$P(\lambda) = \det(A - \lambda I_n)$$

(see (2.3.4) on p.90). In other words, λ is an eigenvalue of A, if and only if the (matrix) operator $A - \lambda I_n$ is singular, or $\lambda \in \sigma(A)$. $\blacksquare$

Example 2 Let $\widetilde{\ell}_2$ be the inner-product space of the one-sided infinite sequences

$$\mathbf{x} = (x_0, x_1, \ldots),$$

with

$$\sum_{j=0}^{\infty} |x_j|^2 < \infty,$$

and with the inner product defined, analogous to ℓ_2, by

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{j=0}^{\infty} x_j \overline{y}_j,$$

where $\mathbf{y} = (y_0, y_1, \ldots) \in \widetilde{\ell}_2$ as well. It is clear that the right-shift operator L on $\widetilde{\ell}_2$, defined by

$$L\mathbf{x} = L(x_0, x_1, \ldots) = (0, x_0, x_1, \ldots), \tag{2.4.2}$$

is a linear operator on $\widetilde{\ell}_2$. Verify that L has no eigenvalues but $0 \in \sigma(L)$.

Solution Let us first observe that $\lambda_0 = 0$ is not an eigenvalue. Indeed, if

$$L\mathbf{x} = 0\mathbf{x} = \mathbf{0},$$

then $(0, x_0, x_1, \ldots) = \mathbf{0}$, or $x_0 = x_1 = \cdots = 0$, or $\mathbf{x} = 0$, so that $\mathbb{N}_{\lambda_0} = \{\mathbf{0}\}$.

Next, if $\lambda_1 \neq 0$ would be an eigenvalue of L, then $L\mathbf{x} = \lambda_1 \mathbf{x}$ can be written as

$$(0, x_0, x_1, x_2, \ldots) = \lambda_1 (x_0, x_1, x_2, x_3, \ldots),$$

or equivalently,

$$0 = \lambda_1 x_0, \ x_0 = \lambda_1 x_1, \ldots, x_j = \lambda_1 x_{j+1}, \ldots.$$

Hence, $\lambda_1 \neq 0$ implies that $x_0 = 0, 0 = x_1, 0 = x_2, \ldots, 0 = x_{j+1}, \ldots$; that is, $\mathbf{x} = 0$. In other words, $\mathbb{N}_{\lambda_1} = \{\mathbf{0}\}$. Therefore, the operator L does not have any eigenvalue.

On the other hand, to show that $0 \in \sigma(L)$, it is equivalent to verifying that

$$L - 0I = L$$

is not invertible on $\widetilde{\ell}_2$. Indeed, if M would be an inverse of L, then M must be a left-shift operator:

$$M(y_0, y_1, y_2, \ldots) = (y_1, y_2, \ldots), \tag{2.4.3}$$

in order to yield

$$ML(x_0, x_2, \ldots) = (x_0, x_1, \ldots),$$

for all $(0, x_0, x_1, \ldots)$ in the range of L.

However, if $y_0 \neq 0$, then $\mathbf{y} = (y_0, y_1, \ldots)$ cannot be recovered from $M\mathbf{y}$ in (2.4.3) since the value y_0 is lost. That is, M does not have an inverse. Hence, L cannot be an inverse of M (or M cannot be an inverse of L). In summary, although L does not have any eigenvalue, the spectrum $\sigma(L)$ of L is not empty. ∎

Definition 2 | **Normal operators and normal matrices** | *A linear operator T on an inner-product space $\mathbb{V}$ over $\mathbb{C}$ or $\mathbb{R}$ said to be normal, if it commutes with its adjoint T^*; that is,*

$$TT^* = T^*T. \tag{2.4.4}$$

In particular, a square matrix A is called a normal matrix if it satisfies

$$AA^* = A^*A.$$

Clearly, if A is Hermitian, namely $A^* = A$, then A is normal.

Note that $\widetilde{\ell}_2$ in Example 2 is an infinite-dimensional vector space. In view of Examples 1–2, we will only discuss finite-dimensional inner-product spaces $\mathbb{V}$ over $\mathbb{F} = \mathbb{C}$ or $\mathbb{R}$ in this section. First we recall the definitions of orthogonal and unitary matrices.

Definition 3 | **Unitary and orthogonal matrices** | *An $n \times n$ matrix A of complex numbers is called a unitary matrix, if*

$$AA^* = I_n. \tag{2.4.5}$$

In particular, if A is a real matrix, then the unitary matrix A is also called an orthogonal matrix; and in this case,

$$AA^T = I_n. \tag{2.4.6}$$

From the definition, we see that a unitary (and particularly an orthogonal) matrix is nonsingular with inverse given by $A^{-1} = A^* = \overline{A}^T$ (and $A^{-1} = A^T$, if A is real), so that $A^*A = AA^* = I_n$.

Theorem 1 *For an $n \times n$ matrix A, write*

$$A = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n] = \begin{bmatrix} \widetilde{\mathbf{a}}_1 \\ \vdots \\ \widetilde{\mathbf{a}}_n \end{bmatrix},$$

where $\mathbf{a}_j$ is the j-th column of A and $\widetilde{\mathbf{a}}_j$ is the j-th row of A. Suppose that A is a unitary matrix. Then

$$\langle \mathbf{a}_j, \mathbf{a}_k \rangle = \delta_{k-j}, \quad \langle \widetilde{\mathbf{a}}_j, \widetilde{\mathbf{a}}_k \rangle = \delta_{k-j}, \ 1 \leq j, k \leq n.$$

The proof follows from the definition (2.4.5) and the formula

$$A^*A = \begin{bmatrix} \bar{\mathbf{a}}_1^T \\ \vdots \\ \bar{\mathbf{a}}_n^T \end{bmatrix} [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n] = \begin{bmatrix} \bar{\mathbf{a}}_1^T \mathbf{a}_1 & \bar{\mathbf{a}}_1^T \mathbf{a}_2 & \ldots & \bar{\mathbf{a}}_1^T \mathbf{a}_n \\ \vdots & \vdots & \vdots & \vdots \\ \bar{\mathbf{a}}_n^T \mathbf{a}_1 & \bar{\mathbf{a}}_n^T \mathbf{a}_2 & \ldots & \bar{\mathbf{a}}_n^T \mathbf{a}_n \end{bmatrix},$$

so that the (j, k)-entry of A^*A is given by $\bar{\mathbf{a}}_j^T \mathbf{a}_k = \overline{\langle \mathbf{a}_j, \mathbf{a}_k \rangle}$, which is also equal to δ_{k-j} by (2.4.5). Similarly, from $AA^* = I_n$, one can obtain $\langle \widetilde{\mathbf{a}}_j, \widetilde{\mathbf{a}}_k \rangle = \delta_{k-j}, 1 \leq j, k \leq n$. ∎

From the above theorem, we see that each row of a unitary matrix $A \in \mathbb{C}^{n,n}$ is a unit vector, and that all the row vectors are orthogonal to one another; that is, the row vectors of a unitary matrix in $\mathbb{C}^{n,n}$ constitute an orthonormal basis of $\mathbb{C}^n$. Similarly, the column vectors of a unitary matrix in $\mathbb{C}^{n,n}$ also constitute an orthonormal basis of $\mathbb{C}^n$. In particular, the set of row vectors and the set of column vectors of an orthogonal matrix in $\mathbb{R}^{n,n}$ are both orthonormal bases of $\mathbb{R}^n$.

Theorem 2 | **Unitary invariance property** | *Every unitary matrix $U \in \mathbb{C}^{n,n}$ has the property:*

$$\langle U\mathbf{x}, U\mathbf{y}\rangle = \langle \mathbf{x}, \mathbf{y}\rangle, \quad \textit{for all} \ \ \mathbf{x}, \mathbf{y} \in \mathbb{C}^n, \tag{2.4.7}$$

or equivalently,

$$\|U\mathbf{x} - U\mathbf{y}\| = \|\mathbf{x} - \mathbf{y}\|, \quad \textit{for all} \ \ \mathbf{x}, \mathbf{y} \in \mathbb{C}^n. \tag{2.4.8}$$

In particular, if $U \in \mathbb{R}^{n,n}$ is an orthogonal matrix, then (2.4.7) *and* (2.4.8) *hold for all* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.

If U is unitary, then

$$\langle U\mathbf{x}, U\mathbf{y}\rangle = \langle \mathbf{x}, U^*U\mathbf{y}\rangle = \langle \mathbf{x}, \mathbf{y}\rangle$$

which is (2.4.7); and

$$\begin{aligned}
\|U\mathbf{x} - U\mathbf{y}\|^2 &= \langle U\mathbf{x} - U\mathbf{y}, U\mathbf{x} - U\mathbf{y}\rangle \\
&= \langle U(\mathbf{x} - \mathbf{y}), U(\mathbf{x} - \mathbf{y})\rangle \\
&= \langle \mathbf{x} - \mathbf{y}, \mathbf{x} - \mathbf{y}\rangle = \|\mathbf{x} - \mathbf{y}\|^2,
\end{aligned}$$

so that (2.4.8) is equivalent to (2.4.7). ∎

The property (2.4.7) can be interpreted as invariance property of "angles" between vectors under unitary transformation, and the property (2.4.8) says that the distance between any two vectors is invariant under unitary transformation as well. In particular, by setting $\mathbf{y} = \mathbf{0}$ in (2.4.8), we see that unitary transformations also preserve vector lengths, in that

$$\|U\mathbf{x}\| = \|\mathbf{x}\|.$$

Of course, these invariant properties are valid for any real vectors under orthogonal transformation.

Theorem 3 $\boxed{\textbf{Spectral theorem}}$ *Let A be an $n \times n$ normal matrix. Then there is a unitary matrix U and a diagonal matrix $\Lambda = \mathrm{diag}\{\lambda_1, \ldots, \lambda_n\}$, such that*

$$A = U\Lambda U^*. \tag{2.4.9}$$

The factorization of the matrix A in (2.4.9) is called the "**spectral decomposition**" of A. The spectral decomposition (2.4.9) can be re-formulated as

$$U^{-1}AU = \mathrm{diag}\{\lambda_1, \ldots, \lambda_n\}. \tag{2.4.10}$$

Hence, any normal matrix A is "diagonizable", in the sense that A is "similar" to a diagonal matrix, namely: the diagonal matrix $\Lambda = \mathrm{diag}\{\lambda_1, \ldots, \lambda_n\}$. Moreover, from the equivalent formulation

$$AU = U\Lambda,$$

it follows that each of $\lambda_1, \ldots, \lambda_n$, on the main diagonal of Λ, is an eigenvalue of A, and the jth column of U is an eigenvector of A associated with the eigenvalue λ_j, for all $j = 1, \ldots, n$.

Examples of normal matrices, as defined by (2.4.4), include all unitary matrices and all Hermitian matrices.

To prove the spectral theorem, let us first establish the following weaker result, where the diagonal matrix Λ is replaced by an upper-triangular matrix $T = [t_{jk}]_{1 \le j, k \le n}$, with $t_{jk} = 0$ for all $j > k$.

Theorem 4   **Unitary-triangular matrix decomposition**   *For any given $n \times n$ matrix A, there exist a unitary matrix U and an upper-triangular matrix T, such that*

$$A = UTU^*. \tag{2.4.11}$$

Proof   This theorem can be proved by mathematical induction, since the theorem trivially holds for $n = 1$. For $n > 1$ and $A \in \mathbb{C}^{n \times n}$, let $\mathbf{v}_1$ be some unit eigenvector associated with an eigenvalue λ_1 of A. We then formulate the unitary matrix

$$V = [\mathbf{v}_1 \ \mathbf{v}_2 \ \cdots \ \mathbf{v}_n],$$

by extending $\mathbf{v}_1$ to an orthonormal basis $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ of $\mathbb{C}^n$. Since $\mathbf{v}_2, \ldots, \mathbf{v}_n$ are orthogonal to $\mathbf{v}_1$, and $A\mathbf{v}_1 = \lambda_1 \mathbf{v}_1$, we have

$$V^*AV = \begin{bmatrix} \mathbf{v}_1^* \\ \vdots \\ \mathbf{v}_n^* \end{bmatrix} [\lambda_1 \mathbf{v}_1 \ * \ \cdots \ *] = \left[\begin{array}{ccccc} \lambda_1 & \vdots & * & \ldots & * \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ O & \vdots & & B & \end{array} \right],$$

where O is a zero column vector and B is an $(n-1) \times (n-1)$ matrix, which, by applying the induction hypothesis, can be written as

$$B = WT_1W^*,$$

where W is some $(n-1) \times (n-1)$ unitary matrix and T_1 an $(n-1) \times (n-1)$ upper-triangular matrix. Therefore, by introducing the matrix

$$U = V \left[\begin{array}{ccc} 1 & \vdots & O \\ \cdots & \cdots & \cdots \\ O & \vdots & W \end{array} \right],$$

we have the following decomposition formulation:

$$
U^*AU = \begin{bmatrix} 1 & \vdots & O \\ \cdots & \cdots & \cdots \\ O & \vdots & W^* \end{bmatrix} \begin{bmatrix} \lambda_1 & \vdots & * & \cdots & * \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ O & \vdots & & B & \end{bmatrix} \begin{bmatrix} 1 & \vdots & O \\ \cdots & \cdots & \cdots \\ O & \vdots & W \end{bmatrix}
$$

$$
= \begin{bmatrix} \lambda_1 & \vdots & * & \cdots & * \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ O & \vdots & & W^*BW & \end{bmatrix} = \begin{bmatrix} \lambda_1 & \vdots & * & \cdots & * \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ O & \vdots & & T_1 & \end{bmatrix},
$$

where the matrix on the right is an upper-triangular matrix. Since U is unitary (because both V and W are unitary), we may conclude that A has the desired matrix decomposition as in (2.4.11), completing the induction argument.   ■

We are now ready to prove Theorem 3, by applying Theorem 4.

Proof of Theorem 3   By Theorem 4, we may write $A = UTU^*$, where U is unitary and $T = [t_{jk}]_{1 \leq j,k \leq n}$ upper-triangular, with $t_{jk} = 0$ for all $j > k$. In the following, we will prove that T is actually a diagonal matrix, and thereby complete the proof of the theorem.

Indeed, from the assumption that A is normal, we have

$$
\begin{aligned}
U(T^*T)U^* &= (UT^*U^*)(UTU^*) = A^*A \\
&= AA^* = (UTU^*)(UT^*U^*) = U(TT^*)U^*,
\end{aligned}
$$

so that

$$
T^*T = TT^*. \tag{2.4.12}
$$

Observe that the $(1, 1)$-entry of T^*T is $|t_{11}|^2$, while the $(1, 1)$-entry of TT^* is $|t_{11}|^2 + |t_{12}|^2 + \cdots + |t_{1n}|^2$. Hence, from (2.4.12), we have

$$
|t_{11}|^2 = |t_{11}|^2 + |t_{12}|^2 + \cdots + |t_{1n}|^2,
$$

and therefore

$$
t_{12} = 0, \ t_{13} = 0, \ \ldots, \ t_{1n} = 0.
$$

This enables us to write

$$
T = \begin{bmatrix} 1 & \vdots & O \\ \cdots & \cdots & \cdots \\ O & \vdots & T_1 \end{bmatrix},
$$

where $T_1 = [t_{jk}]_{2 \leq j,k \leq n}$ is an $(n-1) \times (n-1)$ upper-triangular matrix. By applying (2.4.12) again, we obtain $T_1^*T_1 = T_1T_1^*$, which yields

$$
t_{23} = t_{24} = \ldots = t_{2n} = 0,
$$

by the same argument as above. Hence, repeating the same process again and again, we may conclude that $T = \Lambda$ is a diagonal matrix, so that $A = UTU^* = U\Lambda U^*$, completing the proof of Theorem 3. ∎

Next we derive the following property of operator norms for normal operators.

Theorem 5 | **Norm preservation by adjoint of normal operators** | *Let T be a normal operator on an inner-product space $\mathbb{V}$. Then T satisfies*

$$\|T\mathbf{x}\| = \|T^*\mathbf{x}\|, \quad \text{for any } \mathbf{x} \in \mathbb{V}; \tag{2.4.13}$$

or equivalently, $\|\|T\|\|_{\mathbb{V}} = \|\|T^*\|\|_{\mathbb{V}}$.

Proof The proof of (2.4.13) is evident from the definition of normal operators. Indeed, for any $\mathbf{v} \in \mathbb{V}$, since $T^*T = TT^*$, we have

$$\begin{aligned}
\|T\mathbf{v}\|^2 &= \langle T\mathbf{v}, T\mathbf{v} \rangle = \langle T^*T\mathbf{v}, \mathbf{v} \rangle \\
&= \langle TT^*\mathbf{v}, \mathbf{v} \rangle = \langle T^*\mathbf{v}, T^*\mathbf{v} \rangle = \|T^*\mathbf{v}\|^2.
\end{aligned}$$

Consequently, by the definition of operator norms in (2.2.4) on p.80, we have, by using the notation in (2.2.5) on the same page, that $\|\|T\|\|_{\mathbb{V}} = \|\|T^*\|\|_{\mathbb{V}}$. ∎

Exercises

Exercise 1 Let A be an $n \times m$ and B an $n \times p$ matrix. Denote $A^*B = C = [c_{j,k}]$. Show that if $A = [\mathbf{a}_1, \ldots, \mathbf{a}_n]$ and $B = [\mathbf{b}_1, \ldots, \mathbf{b}_p]$, then $c_{j,k} = \langle \bar{\mathbf{a}}_j, \bar{\mathbf{b}}_k \rangle = \langle \mathbf{a}_j, \mathbf{b}_k \rangle$.

Exercise 2 Let A be an $m \times n$ and B a $p \times n$ matrix. Denote $AB^* = C = [c_{j,k}]$. Show that if $A = \begin{bmatrix} \mathbf{a}_1 \\ \vdots \\ \mathbf{a}_m \end{bmatrix}$ and $B = \begin{bmatrix} \mathbf{b}_1 \\ \vdots \\ \mathbf{b}_p \end{bmatrix}$, then $c_{j,k} = \langle \mathbf{a}_j, \mathbf{b}_k \rangle$.

Exercise 3 For each of the following matrices, compute its eigenvalues and corresponding eigenvectors with unit norm; and compute the required unitary (or orthogonal) matrix to formulate its spectral decomposition.

(a) $A_1 = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$.

(b) $A_2 = \begin{bmatrix} 2 & -4 \\ 1 & 1 \end{bmatrix}$.

(c) $A_3 = \begin{bmatrix} 1 & 0 & 1 \\ 0 & -1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$.

(d) $A_4 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix}$.

Exercise 4 Identify which of the following matrices are diagonalizable. Justify your answers with rigorous arguments.

(a) $B_1 = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}$.

(b) $B_2 = \begin{bmatrix} -1 & 0 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$.

(c) $B_3 = \begin{bmatrix} 0 & 1 & 2 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{bmatrix}$.

(d) $B_4 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$.

Exercise 5 In the following, identify those matrices that are normal, those that are unitary, and those that are Hermitian (that is, self-adjoint). Justify your answers with rigorous arguments.

(a) $C_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -i \\ -i & 1 \end{bmatrix}$.

(b) $C_2 = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$.

(c) $C_3 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -i \\ i & 1 \end{bmatrix}$.

(d) $C_4 = \frac{1}{3} \begin{bmatrix} 1 & -2 & 2 \\ 2 & 2 & 1 \\ 2 & -1 & -2 \end{bmatrix}$.

Exercise 6 Compute the trace of each of the matrices A_1, A_2, A_3, A_4 (in Exercise 3) and C_1, C_2, C_3, C_4 (in Exercise 5)

Exercise 7 In the following, identify all unitary (or orthogonal) matrices. For those that are not unitary (or orthogonal), normalize the appropriate column vectors (by dividing them with their norms, called normalization constants), so that the modified matrices become unitary (or orthogonal).

(a) $D_1 = \begin{bmatrix} 2 & -2 \\ 2 & 2 \end{bmatrix}$.

(b) $D_2 = \begin{bmatrix} 1 & 2 & 2 \\ 2 & -2 & 1 \\ 2 & 1 & -2 \end{bmatrix}$.

(c) $D_3 = \frac{1}{2} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \end{bmatrix}.$

(d) $D_4 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & i & -1 & -i \\ 1 & -i & -1 & i \end{bmatrix}.$

Exercise 8 For each of D_1, D_2, D_3, D_4 in Exercise 7, without performing matrix multiplications, compute the distance between the transformed vectors, by applying (2.4.8) for unitary matrices.

   Hint: If D_j is not unitary but its modification is, as in Exercise 7, then the normalization constants can be transferred to the corresponding components of the given vectors **x** and **y**.

(a) $\|D_1\mathbf{x} - D_1\mathbf{y}\|$, where $\mathbf{x} = (1, 2)$, $\mathbf{y} = (-1, 1)$.
(b) $\|D_2\mathbf{x} - D_2\mathbf{y}\|$, where $\mathbf{x} = (1, -1, 1))$, $\mathbf{y} = (0, 1, 0)$.
(c) $\|D_3\mathbf{x} - D_3\mathbf{y}\|$, where $\mathbf{x} = (1, 0, -1, 1)$, $\mathbf{y} = (0, 1, 1, 1)$.
(d) $\|D_4\mathbf{x} - D_4\mathbf{y}\|$, where $\mathbf{x} = (i, 0, -i, 1)$, $\mathbf{y} = (0, 1, 0, 0)$.

Exercise 9 As a continuation of Exercises 7–8, compute the angles between the two vectors $D_j\mathbf{x}$ and $D_j\mathbf{y}$, for each $j = 1, \ldots, 4$, where **x**, **y** are given in (a), (b), (c), or (d) in Exercise 8

Chapter 3
Spectral Methods and Applications
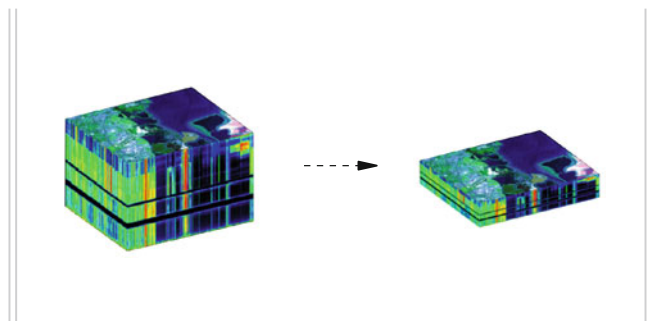
The study of linear transformations, eigenvalue problems, and spectral decomposition in Chap. 2 is extended in this chapter to singular value decomposition, principal component analysis, and various operator norms for two areas of applications to data representations.

In Sect. 3.1, the spectral decomposition studied in Chap. 2 is applied to extending the eigenvalue problem for square matrices to the discussion of the singular value problem for rectangular matrices. As a result, eigenvalues of square matrices are extended to singular values of rectangular matrices, and the spectral decomposition of normal matrices studied in Sect. 2.3 is applied to the derivation of the singular value decomposition (SVD) of arbitrary matrices, particularly those that are not square. It is shown that singular values are preserved under unitary transformations, and this result is applied to introduce the concept of principal components as well as the method of principal component analysis (PCA). When the matrix represents a dataset, in that each row vector is a data point, the method of PCA provides a new coordinate system to facilitate better analysis of the given data. We end this section by proving that the singular values of a square matrix are precisely the absolute values of its eigenvalues, if and only if the matrix is a normal matrix.

In Sect. 3.2, the operator norm introduced in Sect. 2.1 of Chap. 2 is elaborated and new operator norms for matrix transformations are introduced. In particular, when the linear transformation is a matrix $B \in \mathbb{C}^{m,n}$ and the Euclidean norm is used for both $\mathbb{C}^n$ and $\mathbb{C}^m$, then the operator norm $\|\|B\|\|_2$ of B is shown to be the largest singular value of the matrix B. In addition, the Frobenius norm is introduced for quantifying the error of approximation of a given matrix by matrices with an arbitrarily pre-assigned lower rank. More generally, the operator norm $\|\|B\|\|_2$ and Frobenius norm of B are extended, from $p = \infty$ and 2, to the Schattern p-norm for arbitrary $0 \le p \le \infty$, defined by the ℓ^p-measurement of the singular values of B. This more general Schattern norm, particularly for $p = 0$ and $p = 1$, will be instrumental to the study of the subject of compressed (or compressive) sensing in a

forthcoming publication of this book series. It is shown in this section that lower-rank matrix approximation of a given matrix B with rank $= r$ is accomplished by writing B as a sum Σ of rank-1 matrices by re-formulating the SVD of B, while its best approximant with pre-assigned rank not exceeding $d < r$ is unique and is given by the partial sum, consisting of the first d terms of Σ.

The last two sections of this chapter, Sects. 3.3 and 3.4, are devoted to the application of SVD and PCA to the study of two areas in data representations, with the first on least-squares data estimation and solving (or finding the closest solution of) an arbitrary system of linear equations, which could be over-determined or under-determined; and the second on representation of data in a lower-dimensional space. In Sect. 3.3, the notion of pseudo-inverse of a possibly rectangular matrix is introduced in terms of the singular values. Several key properties of the pseudo-inverse are derived, including its agreement with the matrix inverse for non-singular square matrices. The other properties are shown to uniquely determine the definition of the pseudo-inverse. With the pseudo-inverse, the solution (or closest solution) of an arbitrary system of linear equations can be explicitly formulated. In addition, least-squares estimation of a given data set is formulated in terms of the pseudo-inverse, and it is shown that this least-squares solution is unique under the constraint of minimum ℓ^2-norm. In Sect. 3.4, motivation of the need for high-dimensional data is discussed by considering spectral curves of digital images. An explicit formula of the dimension-reduced data is derived by applying SVD and PCA; and a result that describes the best approximation in the ℓ^2-norm is derived with two demonstrative toy-examples.

3.1 Singular Value Decomposition and Principal Component Analysis

In this section, we extend the concept of eigenvalues for square matrices to the notion of singular values that applies even to rectangular matrices in general. We will then introduce and derive the matrix factorization, called singular value decomposition (SVD), which has a host of applications, such as those to be discussed in Sect. 3.3 and Sect. 3.4 of this chapter.

Let B be an $m \times n$ matrix, and consider its corresponding Gram matrix A, defined by

$$A = BB^*, \tag{3.1.1}$$

where $B^* = \left(\overline{B}\right)^T$. Then A is self-adjoint and positive semi-definite (spd) (see Remark 5 on p.102), and is therefore normal. Hence, A has the spectral decomposition

$$A = U\Lambda U^* \tag{3.1.2}$$

by Theorem 3 on p.108, with diagonal matrix $\Lambda = \text{diag}\{\lambda_1, \ldots, \lambda_m\}$ and unitary matrix

$$U = [\mathbf{u}_1 \quad \cdots \quad \mathbf{u}_m],$$

where for each $j = 1, \ldots, m$, $(\lambda_j, \mathbf{u}_j)$ is an eigenvalue-eigenvector pair of the matrix A. Furthermore, since A is spd, we may, and will, write $\lambda_j = \sigma_j^2$, where

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_m = 0 \tag{3.1.3}$$

for some r, with $0 \leq r \leq m$. Hence, the diagonal matrix Λ in the spectral decomposition (3.1.2) has the more precise formulation

$$\Lambda = \text{diag}\{\sigma_1^2, \ldots, \sigma_r^2, 0, \ldots, 0\}, \tag{3.1.4}$$

where we have adopted the standard convention that $\{\sigma_{r+1}, \ldots, \sigma_m\}$ is an empty set if $r = m$.

Furthermore, recall from Theorem 4 on p.76 that

$$\text{rank}(B) = \text{rank}(B\,B^*) = \text{rank}(\Lambda) = r,$$

so that $r \leq \min\{m, n\}$. Let

$$\Sigma_r = \text{diag}\{\sigma_1, \ldots, \sigma_r\} \tag{3.1.5}$$

and consider the $m \times n$ matrix

$$S = \begin{bmatrix} \Sigma_r & \vdots & O \\ \cdots & \cdots & \cdots \\ O & \vdots & O \end{bmatrix}, \tag{3.1.6}$$

where O denotes the zero matrix (possibly with different dimensions), so that

$$S = \begin{bmatrix} \Sigma_n \\ \cdots \\ O \end{bmatrix} \text{ or } S = \begin{bmatrix} \Sigma_m & \vdots & O \end{bmatrix} \tag{3.1.7}$$

if $r = n < m$ or $r = m < n$, respectively. Observe that the diagonal matrix Λ in (3.1.4) can be written as

$$\Lambda = SS^T = SS^*, \tag{3.1.8}$$

and the spectral decomposition of A in (3.1.2) can be re-formulated as

$$A = USS^*U^* = (US)(US)^*. \tag{3.1.9}$$

Since $A = BB^*$, the formulation in (3.1.9) may suggest a factorization of the matrix B as some unitary transformation U of the "diagonal" rectangular matrix S. Unfor-

tunately, this is not quite correct yet, but an additional unitary transformation V of S will do the job. More precisely, we will show the existence of two unitary matrices U and V, of dimensions $m \times m$ and $n \times n$, respectively, such that

$$B = USV^* \qquad (3.1.10)$$

(see Theorem 2 on p.121 to be derived later in this section).

To understand the factorization in (3.1.10), let us write

$$U = [\mathbf{u}_1, \ldots, \mathbf{u}_m], \quad V = [\mathbf{v}_1, \ldots, \mathbf{v}_n], \qquad (3.1.11)$$

where $\{\mathbf{u}_1, \ldots, \mathbf{u}_m\}$ and $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ are orthonormal bases of $\mathbb{C}^m$ and $\mathbb{C}^n$, respectively. Then it follows from (3.1.10) that $BV = US$ and $B^*U = VS^T$, so that

(i) if $n < m$, then

$$B\mathbf{v}_j = \sigma_j \mathbf{u}_j, \quad B^*\mathbf{u}_j = \sigma_j \mathbf{v}_j, \quad \text{for } j = 1, \ldots, n, \qquad (3.1.12)$$
$$B^*\mathbf{u}_j = \mathbf{0}, \quad \text{for } j = n+1, \ldots, m;$$

(ii) if $n \geq m$, then

$$B\mathbf{v}_j = \sigma_j \mathbf{u}_j, \quad B^*\mathbf{u}_j = \sigma_j \mathbf{v}_j, \quad \text{for } j = 1, \ldots, m, \qquad (3.1.13)$$
$$B\mathbf{v}_j = \mathbf{0}, \quad \text{for } j = m+1, \ldots, n$$

(see Exercise 1).

Definition 1 $\boxed{\text{Singular values}}$ *In (3.1.10), the diagonal entries $\sigma_1, \ldots, \sigma_r$ of Σ_r in (3.1.6) are called the (non-zero) singular values of the matrix B, and the pair $(\mathbf{v}_j, \mathbf{u}_j)$ of vectors in (3.1.12) or (3.1.13) is called a singular-vector pair associated with the singular value σ_j.*

Clearly if $(\mathbf{v}_j, \mathbf{u}_j)$ is a singular-vector pair of B associated with σ_j, then $(\mathbf{u}_j, \mathbf{v}_j)$ is a singular-vector pair of B^* associated with the same σ_j.

Example 1 For the matrix

$$B_1 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & -1 \end{bmatrix}$$

with $m = 2$ and $n = 3$, verify that $\sigma_1 = \sqrt{2}$, $\sigma_2 = 1$ are singular values of B_1, with corresponding singular-vector pairs $(\mathbf{v}_1, \mathbf{u}_1)$, $(\mathbf{v}_2, \mathbf{u}_2)$, and $B_1\mathbf{v}_3 = \mathbf{0}$, where

$$\mathbf{v}_1 = \left[\tfrac{1}{\sqrt{2}} \; 0 \; -\tfrac{1}{\sqrt{2}} \right]^T,$$

$$\mathbf{v}_2 = \left[0 \; 1 \; 0 \right]^T,$$

$$\mathbf{v}_3 = \left[\tfrac{1}{\sqrt{2}} \; 0 \; \tfrac{1}{\sqrt{2}} \right]^T,$$

$$\mathbf{u}_1 = \left[0 \; 1 \right]^T \quad \text{and} \quad \mathbf{u}_2 = \left[1 \; 0 \right]^T.$$

Solution

(i) For $\sigma_1 = \sqrt{2}$,

$$B_1 \mathbf{v}_1 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} \tfrac{1}{\sqrt{2}} \\ 0 \\ -\tfrac{1}{\sqrt{2}} \end{bmatrix} = \begin{bmatrix} 0 \\ \sqrt{2} \end{bmatrix}$$

$$= \sqrt{2} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \sigma_1 \mathbf{u}_1;$$

$$B_1^* \mathbf{u}_1 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$$

$$= \sqrt{2} \begin{bmatrix} \tfrac{1}{\sqrt{2}} \\ 0 \\ -\tfrac{1}{\sqrt{2}} \end{bmatrix} = \sigma_1 \mathbf{v}_1;$$

(ii) for $\sigma_2 = 1$,

$$B_1 \mathbf{v}_2 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \sigma_2 \mathbf{u}_2;$$

$$B_1^* \mathbf{u}_2 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = \sigma_2 \mathbf{v}_2;$$

(iii) for $\mathbf{v}_3$,

$$B_1 \mathbf{v}_3 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} \tfrac{1}{\sqrt{2}} \\ 0 \\ \tfrac{1}{\sqrt{2}} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \mathbf{0}.$$

Observe that $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ is an orthonormal basis of $\mathbb{R}^3$ (and of $\mathbb{C}^3$), and $\{\mathbf{u}_1, \mathbf{u}_2\}$ is an orthonormal basis of $\mathbb{R}^2$ (and of $\mathbb{C}^2$). In addition, in this example $2 = m < n = 3$, and (3.1.14) holds with $B = B_1$. ∎

Theorem 1 $\boxed{\textbf{Reduced SVD theorem}}$ *Let B be an $m \times n$ matrix with* $\mathrm{rank}(B) = r$. *Then there exists an $m \times r$ matrix U_1 and an $n \times r$ matrix V_1, with*

$$U_1^* U_1 = I_r, \quad V_1^* V_1 = I_r, \tag{3.1.14}$$

such that B has the reduced SVD

$$B = U_1 \Sigma_r V_1^*, \tag{3.1.15}$$

where $\Sigma_r = \mathrm{diag}\{\sigma_1, \ldots, \sigma_r\}$ with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$. Furthermore, if B is a real-valued matrix, then U_1 and V_1 in (3.1.15) can be chosen to be real-valued matrices.

Proof To prove Theorem 1, let $A = BB^*$. Then A has the spectral decomposition (3.1.2) for some $m \times m$ unitary matrix U and $\Lambda = \mathrm{diag}\{\sigma_1^2, \ldots, \sigma_m^2\}$ with

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_m = 0,$$

where $0 \leq r \leq \min\{m, n\}$. Furthermore, by Theorem 4 on p.76, we have $\mathrm{rank}(B) = \mathrm{rank}(BB^*)$, so that $\mathrm{rank}(B) = \mathrm{rank}(A) = \mathrm{rank}(\Lambda) = r$.

Write U as $U = [U_1 \,\vdots\, U_2]$, with U_1 being the $m \times r$ matrix consisting of the first r columns of U. Then $U_1^* U_1 = I_r$ and $U_1^* U_2 = O$, where I_r denotes, as usual, the $r \times r$ identity matrix. Observe from (3.1.2), that

$$U_1^* B B^* U_1 = U_1^* U \Lambda U^* U_1 = [I_r \ O] \Lambda [I_r \ O]^* = (\Sigma_r)^2. \tag{3.1.16}$$

Similarly, it can be shown that $U_2^* B B^* U_2 = O$, yielding

$$U_2^* B = O. \tag{3.1.17}$$

Next, we define V_1 by

$$V_1 = B^* U_1 \Sigma_r^{-1}. \tag{3.1.18}$$

Then V_1 satisfies (3.1.15). Indeed, from (3.1.17) and the definition of V_1 in (3.1.18), we have

$$U^*(B - U_1 \Sigma_r V_1^*) = \begin{bmatrix} U_1^* B \\ U_2^* B \end{bmatrix} - \begin{bmatrix} I_r \\ O \end{bmatrix} \Sigma_r V_1^*$$
$$= \begin{bmatrix} U_1^* B - \Sigma_r V_1^* \\ O \end{bmatrix} = O.$$

Thus, since U^* is nonsingular, we have $B - U_1 \Sigma_r V_1^* = O$; that is, $B = U_1 \Sigma_r V_1^*$, as desired.

Furthermore, for real matrices B, the unitary matrix U in the spectral decomposition of $A = BB^* = BB^T$ in (3.1.2) can be chosen to be an $m \times m$ orthogonal matrix, and hence the matrix V_1 defined by (3.1.18) is also real. ∎

Theorem 2 $\boxed{\textbf{Full SVD theorem}}$ *Let B be an $m \times n$ matrix with* $\mathrm{rank}(B) = r$. *Then there exist $m \times m$ and $n \times n$ unitary matrices U and V, respectively, such that*

$$B = USV^*, \tag{3.1.19}$$

where S is an $m \times n$ matrix given by (3.1.5) *for $\Sigma_r = \mathrm{diag}\{\sigma_1, \ldots, \sigma_r\}$ with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$. Furthermore, if B is a real matrix, then the unitary matrices U and V in* (3.1.19) *can be chosen to be orthogonal matrices.*

Proof To prove Theorem 2, let us again write $A = BB^*$ and consider the matrix V_1 defined by (3.1.18) so that (3.1.15) holds. Also let $U = [U_1 \ \vdots \ U_2]$, where U_1 and U_2 are the matrices introduced in the proof of Theorem 1. Then by (3.1.18) and (3.1.16), we have

$$V_1^* V_1 = \Sigma_r^{-1} U_1^* BB^* U_1 \Sigma_r^{-1} = \Sigma_r^{-1} (\Sigma_r)^2 \Sigma_r^{-1} = I_r.$$

Hence, the columns of V_1 constitute an orthonormal family in $\mathbb{C}^n$, and we may extend V_1 to a unitary matrix $V = [V_1 \vdots V_2]$ by introducing another matrix V_2 with orthonormal column vectors. Thus, it follows from (3.1.15) that

$$B = U_1 \Sigma_r V_1^* = U_1 [\Sigma_r \ O][V_1 \ V_2]^* = [U_1 \ U_2] \begin{bmatrix} \Sigma_r & O \\ O & O \end{bmatrix} V^* = USV^*,$$

completing the proof of (3.1.19).

Furthermore, for real matrices B, the unitary matrix U in the spectral decomposition (3.1.2) of $A = BB^T$ can be chosen to be an orthogonal matrix. In addition, as already shown above, the columns of V_1 defined by (3.1.18) constitute an orthonormal family of $\mathbb{R}^n$, so that V_1 can be extended to an orthogonal matrix $V \in \mathbb{R}^{n,n}$ that satisfies $B = USV^T$. ∎

From a full SVD of B in (3.1.19), the construction of a reduced SVD (3.1.15) of B is obvious, simply by keeping only the first r columns of U and V to obtain U_1 and V_1, respectively. Conversely, from a reduced SVD (3.1.15) of B, it is also possible to recover a full SVD of B in (3.1.19) by extending Σ_r to S, defined in (3.1.5), as well as extending U_1 and V_1 to unitary matrices U and V, respectively. In the literature, both the reduced SVD (3.1.15) and full SVD (3.1.19) are called SVD of B.

Remark 1 To compute the singular value decomposition (SVD) of a rectangular matrix B, the first step is to compute the eigenvalues of BB^*. Then the non-zero singular values of B are the positive square-roots of the non-zero eigenvalues of BB^*. The unitary matrix U in the SVD of B is the unitary matrix in the spectral

decomposition of BB^*. It is important to emphasize that eigenvectors of BB^* associated with the same eigenvalue must be orthogonalized by applying the Gram-Schmidt process, and that all eigenvectors must be normalized to have unit norm. Of course U and V are not unique, although the singular values are unique. ∎

Remark 2 To compute the singular value decomposition (SVD) of a rectangular matrix B of dimension $m \times n$ with $n < m$, the computational cost can be reduced by computing the spectral decomposition of the $n \times n$ matrix B^*B instead of BB^*, which has larger dimension. To do so, simply replace B by B^* and consider $A = (B^*)(B^*)^*$. Hence, the reduced SVD and full SVD are given by $B^* = U_1 \Sigma_r V_1^*$ and $B^* = U\Lambda V^*$, respectively; so that

$$B = V_1 \Sigma_r U_1^* = V\Lambda U^*.$$

∎

Example 2 Let $B = B_1$ in Example 1; that is,

$$B = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & -1 \end{bmatrix}.$$

Compute the SVD of B.

Solution Since B^*B is 3×3 and BB^* is 2×2, we compute the eigenvalues of one with lower dimension, namely:

$$BB^* = BB^T = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 0 & -1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix},$$

by taking the determinant of $\begin{bmatrix} 1-\lambda & 0 \\ 0 & 2-\lambda \end{bmatrix}$, yielding the eigenvalues $\sigma_1^2 = 2$ and $\sigma_2^2 = 1$ (since they are arranged in decreasing order). Then the (non-zero) singular values of B are:

$$\sigma_1 = \sqrt{2}, \ \sigma_2 = 1.$$

To compute $\mathbf{u}_1$ and $\mathbf{u}_2$, note that

$$BB^* - \sigma_j^2 I_2 = \begin{bmatrix} 1-\sigma_j^2 & 0 \\ 0 & 2-\sigma_j^2 \end{bmatrix}, \ j = 1, 2;$$

that is, $\begin{bmatrix} -1 & 0 \\ 0 & 0 \end{bmatrix}$ and $\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$. Hence, we may select

$$\mathbf{u}_1 = [0 \ 1]^T \ \text{ and } \ \mathbf{u}_2 = [1 \ 0]^T,$$

yielding

$$U = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Since $r = m = 2$, in this case $U_1 = U$, and V_1 as defined by (3.1.18) is simply

$$V_1 = B^*U_1\Sigma_2^{-1} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 \\ 0 & 1 \\ -\frac{1}{\sqrt{2}} & 0 \end{bmatrix}.$$

Hence, the reduced SVD of B is given by

$$B = U_1\Sigma_2 V_1^* = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \sqrt{2} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}} \\ 0 & 1 & 0 \end{bmatrix}.$$

To obtain a full SVD of B, V can be obtained by extending V_1 to a 3×3 orthogonal matrix:

$$V = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ 0 & 1 & 0 \\ -\frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \end{bmatrix}.$$

Therefore the SVD of B is given by

$$B = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \sqrt{2} & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}} \\ 0 & 1 & 0 \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \end{bmatrix}.$$

∎

Example 3 Compute the SVD of B, where

$$B = \begin{bmatrix} 1 & 0 & i \\ 1 & -i & 0 \\ -1 & 0 & -1 \\ 0 & 1 & i \end{bmatrix}.$$

Solution Since B is 4×3 with $m = 4 > n = 3$, we consider the SVD of B^* first and compute the spectral decomposition of $A = (B^*)(B^*)^* = B^*B$:

$$A = \begin{bmatrix} 1 & 1 & -1 & 0 \\ 0 & i & 0 & 1 \\ -i & 0 & -1 & -i \end{bmatrix} \begin{bmatrix} 1 & 0 & i \\ 1 & -i & 0 \\ -1 & 0 & -1 \\ 0 & 1 & i \end{bmatrix} = \begin{bmatrix} 3 & -i & 1+i \\ i & 2 & i \\ 1-i & -i & 3 \end{bmatrix}.$$

To compute the eigenvalues of A, we evaluate the determinant of the matrix $\lambda I_3 - A$ and factorize the characteristic polynomial, yielding

$$(\lambda - 2)(\lambda^2 - 6\lambda + 5) = (\lambda - 2)(\lambda - 5)(\lambda - 1),$$

so that $\lambda_1 = 5$, $\lambda_2 = 2$, $\lambda_3 = 1$, when arranged in the decreasing order. Hence, the (non-zero) singular values of B are

$$\sigma_1 = \sqrt{5}, \ \sigma_2 = \sqrt{2}, \ \sigma_3 = 1.$$

To compute the eigenvectors, we simply solve the three homogeneous linear systems $(A - \lambda_j I_3)\mathbf{x} = \mathbf{0}$, $j = 1, 2, 3$. After dividing each solution by its Euclidean norm, we obtain the normalized eigenvectors $\mathbf{u}_j$ associated with λ_j, for $j = 1, 2, 3$, listed as follows:

$$\mathbf{u}_1 = \frac{1}{2\sqrt{6}} \begin{bmatrix} 1 - 3i \\ 2 \\ -1 - 3i \end{bmatrix}, \ \mathbf{u}_2 = \frac{1}{\sqrt{3}} \begin{bmatrix} 1 \\ -1 \\ -1 \end{bmatrix}, \ \mathbf{u}_3 = \frac{1}{2} \begin{bmatrix} 1 \\ 1 - i \\ i \end{bmatrix}.$$

Let $U = [\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3]$. Since $r = 3$, we have $U_1 = U$ and $\Sigma_3 = \mathrm{diag}\{\sqrt{5}, \sqrt{2}, 1\}$. Applying (3.1.18), we have

$$
\begin{aligned}
V_1 &= (B^*)^* U_1 \Sigma_3^{-1} = BU\Sigma_3^{-1} \\[2mm]
&= \begin{bmatrix} 1 & 0 & i \\ 1 & -i & 0 \\ -1 & 0 & -1 \\ 0 & 1 & i \end{bmatrix} \begin{bmatrix} \frac{1-3i}{2\sqrt{6}} & \frac{1}{\sqrt{3}} & \frac{1}{2} \\[1mm] \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{3}} & \frac{1-i}{2} \\[1mm] \frac{-1-3i}{2\sqrt{6}} & -\frac{1}{\sqrt{3}} & \frac{i}{2} \end{bmatrix} \mathrm{diag}\left\{ \frac{1}{\sqrt{5}}, \frac{1}{\sqrt{2}}, 1 \right\} \\[2mm]
&= \begin{bmatrix} \frac{2-2i}{\sqrt{30}} & \frac{1-i}{\sqrt{6}} & 0 \\[2mm] \frac{1-5i}{2\sqrt{30}} & \frac{1+i}{\sqrt{6}} & -\frac{i}{2} \\[2mm] \frac{3i}{\sqrt{30}} & 0 & -\frac{1+i}{2} \\[2mm] \frac{5-i}{2\sqrt{30}} & -\frac{1+i}{\sqrt{6}} & -\frac{i}{2} \end{bmatrix}.
\end{aligned}
$$

Thus, the reduced SVD for B^* is given by $B^* = U\Sigma_3 V_1^*$, from which we arrive at the following reduced SVD for B:

$$B = V_1 \Sigma_3 U^*.$$

To obtain the full SVD for B, we must extend

$$V_1 = [\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3] \in \mathbb{C}^{4,3}$$

to a unitary matrix

$$V = [\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4] \in \mathbb{C}^{4,4}.$$

To do so, we may select any vector $\mathbf{w}_4 \in \mathbb{C}^4$ which is not a linear combination of $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ and apply the Gram-Schmidt orthogonalization process, introduced in Sect. 1.4 of Chap. 1, to the linearly independent set $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{w}_4$. In this example, we may choose $\mathbf{w}_4 = [1, 0, 0, 0]^T$, and compute:

$$
\begin{aligned}
\widetilde{\mathbf{w}}_4 &= \mathbf{w}_4 - \langle \mathbf{w}_4, \mathbf{v}_1 \rangle \mathbf{v}_1 - \langle \mathbf{w}_4, \mathbf{v}_2 \rangle \mathbf{v}_2 - \langle \mathbf{w}_4, \mathbf{v}_3 \rangle \mathbf{v}_3 \\
&= \mathbf{w}_4 - \frac{2 + 2i}{\sqrt{30}} \mathbf{v}_1 - \frac{1 + i}{\sqrt{6}} \mathbf{v}_2 - 0 \mathbf{v}_3 \\
&= \frac{1}{5} [2, -1 - i, 1 - i, -1 + i]^T,
\end{aligned}
$$

followed by normalization of $\widetilde{\mathbf{w}}_4$:

$$\mathbf{v}_4 = \frac{\widetilde{\mathbf{w}}_4}{\|\widetilde{\mathbf{w}}_4\|} = \frac{1}{\sqrt{20}} [2, -1 - i, 1 - i, -1 + i]^T.$$

With this unitary matrix $V = [\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4]$, we obtain the full SVD $B^* = U [\Sigma_3, O] V^*$ of B^*, yielding the full SVD

$$B = V \begin{bmatrix} \Sigma_3 \\ O \end{bmatrix} U^*,$$

after taking complex conjugation of the transpose.   ∎

In the next two theorems we provide some properties of singular values.

Theorem 3 *Let $B \in \mathbb{C}^{m,n}$. Then the matrices $B^T, \overline{B}, B^*$ have the same singular values as B.*

Proof We first show that B^T and B have the same (non-zero) singular values. Indeed, let $B = USV^*$ be the full SVD for B, where U and V are unitary matrices, and S is an $m \times n$ matrix given by (3.1.6) with $\Sigma_r = \mathrm{diag}\{\sigma_1, \ldots, \sigma_r\}$ and where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ are the non-zero singular values of B. Thus,

$$B^T = \overline{V} S^T \overline{U}^*,$$

where S^T is an $n \times m$ matrix given by

$$S^T = \begin{bmatrix} \Sigma_r^T & O \\ O & O \end{bmatrix} = \begin{bmatrix} \Sigma_r & O \\ O & O \end{bmatrix}.$$

Since $\overline{V}$ and $\overline{U}$ are unitary matrices, $B^T = \overline{V} S^T \overline{U}^*$ is an SVD of B^T. Therefore, the nonzero entries on the main diagonal of Σ_r in S^T are the singular values of B^T. That is, B^T and B have the same non-zero singular values: $\sigma_1, \sigma_2, \ldots, \sigma_r$.

It is easy to see that $\overline{B}$ and B have the same singular values. Since $B^* = (\overline{B})^T$, B^* has the same singular values as $\overline{B}$, which has the same singular values as B. Thus, B^* and B have the same singular values.   ∎

Theorem 4 | **Unitary transformations preserve singular values** | *Let $B \in \mathbb{C}^{m,n}$, and W and R be unitary matrices of dimensions m and n, respectively. Then the singular values of WBR are the same as the singular values of B. More precisely, if $r = \mathrm{rank}(B)$, then $\mathrm{rank}(WBR) = \mathrm{rank}(B) = r$ and*

$$\sigma_j(WBR) = \sigma_j(B), \quad j = 1, \ldots, r,$$

where $\sigma_j(WBR)$ and $\sigma_j(B)$ denote the singular values of WBR and B, respectively, listed in non-increasing order.

Proof The proof of the above theorem follows from the observation that

$$(WBR)(WBR)^* = WBRR^*B^*W^* = WBB^*W^*,$$

which implies that $(WBR)(WBR)^*$ and BB^* are similar matrices, and hence have the same eigenvalues, $\lambda_1 \geq \cdots \geq \lambda_m \geq 0$. Therefore, we may conclude that $\sigma_j(WBR) = \sigma_j(B) = \sqrt{\lambda_j}$, for all $j = 1, \ldots, r$.   ∎

In the next section, Sect. 3.2, we will show that the SVD of a matrix B can be applied to derive a lower-rank matrix approximation of B. But first let us introduce the concept of principal components, as follows.

Definition 2 | **Principal components** | *Let $B = USV^*$ be the SVD of an $m \times n$ matrix B in Theorem 2, where S is given by (3.1.5) with singular values $\sigma_1, \ldots, \sigma_r$ of B in the sub-block Σ_r of S, arranged in non-increasing order as in (3.1.4). Then the singular vector pair $(\mathbf{v}_1, \mathbf{u}_1)$ associated with the largest singular value σ_1 is called the principal component of B. Furthermore, the singular vector pairs $(\mathbf{v}_2, \mathbf{u}_2), \ldots, (\mathbf{v}_r, \mathbf{u}_r)$, associated with the corresponding singular values $\sigma_2, \ldots, \sigma_r$, are called the second, $\ldots, r^{\text{th}}$ principal components of B, respectively.*

Remark 3 | **Uniqueness of principal components** | *Let $B = U_1 \Sigma_r V_1^*$ be the reduced SVD of an $m \times n$ matrix B in Theorem 1, where $\Sigma_r = \mathrm{diag}\{\sigma_1, \ldots, \sigma_r\}$ with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ being the singular values of B. Since $U_1^* U_1 = I_r$, $V_1^* V_1 = I_r$, we have*

$$BB^* = U_1 \Sigma_r^2 U_1^*, \quad B^*B = V_1 \Sigma_r^2 V_1^*,$$

and hence,

$$BB^* U_1 = U_1 \Sigma_r^2, \quad B^*B V_1 = V_1 \Sigma_r^2.$$

Thus, for each $j = 1, \ldots, r$, $\mathbf{u}_j$ is an eigenvector of BB^* associated with eigenvalue σ_j^2 and $\mathbf{v}_j$ is an eigenvector of B^*B associated with σ_j^2. Therefore, if $\sigma_1 > \sigma_2 > \cdots > \sigma_s$ for s, $2 \leq s \leq r$, then σ_j^2, $1 \leq j \leq s$ are simple eigenvalues of BB^* and B^*B as well, and hence, the normalized corresponding eigenvectors $\mathbf{u}_j$ and $\mathbf{v}_j$ are unique. ∎

Remark 4 In application to data analysis, when the matrix $B \in \mathbb{C}^{m,n}$ is a data matrix, with the data in $\mathbb{C}^n$ given by the m rows $\mathbf{b}_1, \ldots, \mathbf{b}_m$ of B, then the principal components of B provide a new "coordinate system", with origin given by the average

$$\mathbf{b}^{\text{av}} = \frac{1}{m} \sum_{j=1}^{m} \mathbf{b}_j.$$

This coordinate system facilitates the analysis of the data, called **principal component analysis (PCA)**, to be discussed in some details in Sect. 3.3. ∎

Recall that singular values are always non-negative by definition. In the following, let us investigate the relationship between absolute values of eigenvalues and singular values for square matrices. More precisely, let B be an $n \times n$ (square) matrix, with singular values $\sigma_1, \ldots, \sigma_n$ (including zero, if any) and eigenvalues $\lambda_1, \ldots, \lambda_n$, arranged in the order: $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$ and $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|$. A natural question to ask is whether or not

$$\sigma_1 = |\lambda_1|, \quad \ldots, \quad \sigma_n = |\lambda_n|? \tag{3.1.20}$$

In general the answer to the question (3.1.20) is negative, as illustrated by the matrices

$$B_1 = \begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix}, \quad B_2 = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix},$$

for which we have

$$B_1 B_1^* = \begin{bmatrix} 5 & 1 \\ 1 & 1 \end{bmatrix}, \quad B_2 B_2^* = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix},$$

with eigenvalues $3 + \sqrt{5}, 3 - \sqrt{5}$ for $B_1 B_1^*$, and eigenvalues $2, 2$ for $B_2 B_2^*$. Hence, the singular values of B_1 are

$$\sigma_1 = \sqrt{3 + \sqrt{5}} = \frac{\sqrt{5} + 1}{\sqrt{2}}, \quad \sigma_2 = \sqrt{3 - \sqrt{5}} = \frac{\sqrt{5} - 1}{\sqrt{2}},$$

while the singular values of B_2 are $\sqrt{2}, \sqrt{2}$. On the other hand, the eigenvalues of B_1 are $2, 1$, while the eigenvalues of B_2 are $1 + i, 1 - i$. It is easy to verify that both B_1 and B_2 provide negative answers to (3.1.20).

On the other hand, consider the following example.

Example 4 Compare the singular values and (absolute values of) the eigenvalues for the two matrices

$$B = \begin{bmatrix} 1 & 1 & 2 \\ 2 & -1 & 1 \\ 1 & 0 & -1 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 0 & -1 \\ 2 & -1 & 1 \end{bmatrix}.$$

Solution First, the determinants of the matrices $\lambda I_3 - B$ and $\lambda I_3 - C$ are given by

$$(\lambda + 2)(\lambda^2 - \lambda - 3), \quad (\lambda - 3)(\lambda + 2)(\lambda - 1),$$

respectively. Thus, the eigenvalues of B are

$$\lambda_1 = \frac{1 + \sqrt{13}}{2}, \ \lambda_2 = -2, \ \lambda_3 = \frac{1 - \sqrt{13}}{2},$$

and the eigenvalues of C are $3, -2, 1$.

On the other hand, since

$$B^T B = C^T C = \begin{bmatrix} 6 & -1 & 3 \\ -1 & 2 & 1 \\ 3 & 1 & 6 \end{bmatrix},$$

the two matrices B and C have the same singular values:

$$\sigma_1 = 3, \ \sigma_2 = 2, \ \sigma_1 = 1.$$

Therefore, the answer to the question (3.1.20) is positive for the matrix C, but negative for the matrix B. ∎

But what is the difference between the two matrices in Example 4? Observe that C is the matrix resulted by exchanging the second and third rows of B; and so we have $C = UB$, where

$$U = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

is unitary. Hence, in view of Theorem 4, the matrices B and C have the same singular values, as demonstrated by the above example. However, while singular values are preserved under unitary transformations, eigenvalues do not enjoy this invariance property.

Also observe that the matrix C in Example 4 is symmetric; and being real, C is self-adjoint and is therefore a normal matrix. This provides an example for the following theorem.

Theorem 5 $\boxed{\textbf{Eigenvalues "=" singular values for normal matrix}}$ *Let $B \in \mathbb{C}^{n,n}$ be any (square) matrix with eigenvalues: $\lambda_1, \ldots, \lambda_n$, and singular values: $\sigma_1, \ldots, \sigma_n$, where multiplicities are listed and magnitudes are arranged in non-increasing order. Then B has the property*

$$|\lambda_j| = \sigma_j, \ \ j = 1, 2, \ldots, n,$$

if and only if B is a normal matrix.

Proof To prove this theorem, let us first consider normal matrices B and apply the spectral decomposition (2.4.9) of Theorem 3 on p.108 to write $B = U\Lambda U^*$, with unitary matrix U and $\Lambda = \mathrm{diag}\{\lambda_1, \ldots, \lambda_n\}$, where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of B, with multiplicities listed and arranged with magnitudes in non-increasing order; that is,

$$|\lambda_1| \geq |\lambda_2| \geq |\lambda_r| > |\lambda_{r+1}| = \cdots = |\lambda_n| = 0,$$

where r denotes the rank of B. For each $j = 1, \ldots, r$, write $\lambda_j = |\lambda_j|e^{i\theta_j}$ for some $\theta_j \in [0, 2\pi)$. Then since U is unitary and the product of two unitary matrices remains to be unitary, the matrix

$$V = U\,\mathrm{diag}\{e^{-i\theta_1}, \ldots, e^{-i\theta_r}, 1, \ldots, 1\}$$

is a unitary matrix. Hence,

$$B = U\,\mathrm{diag}\{|\lambda_1|, \ldots, |\lambda_r|, 0, \ldots, 0\}V^*,$$

is an SVD for B, with singular values $|\lambda_1|, \ldots, |\lambda_r|$.

   To prove the converse, assume that the eigenvalues $\lambda_1, \ldots, \lambda_n$ and singular values $\sigma_1, \ldots, \sigma_n$ of B satisfy $|\lambda_j| = \sigma_j$, for $j = 1, \ldots, n$. Then by Theorem 4 on p.109, the matrix B has the decomposition $B = UTU^*$, where U is unitary and T is upper-triangular. It is evident that the diagonal entries t_{jj} of T are the eigenvalues of B, so that

$$\sum_{j=1}^{n} |t_{jj}|^2 = \sum_{j=1}^{n} |\lambda_j|^2.$$

On the other hand, since σ_j^2 are the eigenvalues of BB^*, it follows from Theorem 1 on p.90 that

$$\sum_{j} \sigma_j^2 = \mathrm{Tr}(BB^*) = \mathrm{Tr}(TT^*) = \sum_{j=1}^{n}\left(\sum_{\ell=1}^{n} t_{j\ell}\overline{t_{j\ell}}\right) = \sum_{1 \leq j, \ell \leq n} |t_{j\ell}|^2,$$

where the second equality follows from the fact that BB^* is similar to TT^*. Thus, from the assumption $|\lambda_j| = \sigma_j$, we have

$$\sum_{j=1}^{n} |t_{jj}|^2 = \sum_{1 \le j, \ell \le n} |t_{j\ell}|^2,$$

which implies that $t_{j\ell} = 0$ for all $j \ne \ell$. Hence, T is a diagonal matrix, so that $TT^* = T^*T$; and therefore,

$$BB^* = UTT^*U^* = UT^*TU = B^*B.$$

That is, B is normal, as we intended to prove.   ∎

<div align="center">

Exercises

</div>

Exercise 1 Let B be an $m \times n$ matrix which can be represented as $B = USV^*$ in (3.1.10), where S is defined in (3.1.6), and U, V be given by (3.1.11), with $\{\mathbf{u}_1, \ldots, \mathbf{u}_m\}$, $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ being orthonormal bases of $\mathbb{C}^m, \mathbb{C}^n$, respectively. Verify the validity of the statements (i) and (ii), as formulated in (3.1.12) and (3.1.13), respectively.

Exercise 2 For the rectangular matrix

$$B = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 0 & 2 \end{bmatrix},$$

verify that $\sigma_1 = 2, \sigma_2 = \sqrt{2}$ are singular values of B, with corresponding singular-vector pairs $(\mathbf{v}_1, \mathbf{u}_1), (\mathbf{v}_2, \mathbf{u}_2)$, and that $B\mathbf{v}_3 = 0$, where

$$\begin{aligned}
\mathbf{v}_1 &= \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}^T, \\
\mathbf{v}_2 &= \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \end{bmatrix}^T, \\
\mathbf{v}_3 &= \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} & 0 \end{bmatrix}^T, \\
\mathbf{u}_1 &= \begin{bmatrix} 0 & 1 \end{bmatrix}^T, \\
\mathbf{u}_2 &= \begin{bmatrix} 1 & 0 \end{bmatrix}^T.
\end{aligned}$$

Exercise 3 Compute the SVD of the following matrices:

(a) $X_1 = \begin{bmatrix} -1 & 1 & 0 \\ 0 & 1 & 0 \end{bmatrix}$,

(b) $X_2 = \begin{bmatrix} 1 & 0 & 1 \\ -1 & 0 & 0 \end{bmatrix}$,

(c) $X_3 = \begin{bmatrix} 0 & 0 & 1 \\ 1 & -1 & 0 \end{bmatrix}$,

(d) $X_4 = \begin{bmatrix} 2 & 1 & -2 \\ -2 & 1 & 2 \end{bmatrix}$.

Exercise 4 Apply Theorem 1 to compute the reduced SVD of the following matrices:

(a) $B_1 = \begin{bmatrix} -1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -1 & 2 & 0 & 0 \end{bmatrix}$,

(b) $B_2 = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 1 \\ -1 & 0 & 0 \end{bmatrix}$,

(c) $B_3 = \begin{bmatrix} 1 & -1 & 0 \\ 1 & -1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$.

Exercise 5 Fill in the details in the derivation of (3.1.17) in the proof of Theorem 1.

Exercise 6 Suppose that B is an $m \times n$ matrix with rank $= r$. Let

$$B^*B = U \Lambda U^*$$

be the spectral decomposition of B^*B, where Λ is the $n \times n$ diagonal matrix $\text{diag}\{\sigma_1^2, \sigma_2^2, \ldots, \sigma_r^2, 0, \ldots, 0\}$. Denote $U = [U_1 \ \vdots \ U_2]$ with U_1 consisting of the first r columns of U. Let V_1 be defined by (3.1.18). Show that

$$\text{(a) } V_1^*V_1 = I_r; \qquad \text{(b) } B = V_1 \Sigma_r U_1^*.$$

Hint: Apply $U_1^*U = [I_r \ \vdots \ O]$ to show that $U_1^*B^*BU_1 = \Sigma_r^2$.

Exercise 7 In Theorem 1, show that if $B \in \mathbb{R}^{m,n}$ is real, then in the reduced SVD (3.1.15) of B, the matrices U_1 and V_1 can be chosen to be real matrices also.

Exercise 8 In Theorem 2, show that if $B \in \mathbb{R}^{m,n}$ is real, then in the full SVD (3.1.19) of B, the unitary matrices U and V can be chosen to be orthogonal matrices (that is, real unitary matrices).

Exercise 9 Compute the singular values of the matrices X, X^T, X^*, and $\overline{X}$ individually, where

$$X = \begin{bmatrix} 1 & i & 0 \\ i & 0 & -1 \end{bmatrix}.$$

Observe that they have the same singular values, as assured by Theorem 3.

Exercise 10 Let X_1, X_2, X_3 be the matrices in Exercise 3, and

$$A = \begin{bmatrix} 3 & 4 \\ -4 & 3 \end{bmatrix}.$$

Apply Theorem 4 and the solution of Exercise 3 to compute the singular values of the following matrices:

(a) $B_1 = AX_1 = \begin{bmatrix} 3 & 4 \\ -4 & 3 \end{bmatrix} \begin{bmatrix} -1 & 1 & 0 \\ 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} -3 & 7 & 0 \\ 4 & -1 & 0 \end{bmatrix}$,

(b) $B_2 = AX_2 = \begin{bmatrix} 3 & 4 \\ -4 & 3 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 \\ -1 & 0 & 0 \end{bmatrix} = \begin{bmatrix} -1 & 0 & 3 \\ -7 & 0 & -4 \end{bmatrix}$,

(c) $B_3 = AX_3 = \begin{bmatrix} 3 & 4 \\ -4 & 3 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 \\ 1 & -1 & 0 \end{bmatrix} = \begin{bmatrix} 4 & -4 & 3 \\ 3 & -3 & -4 \end{bmatrix}$.

Hint: Observe that the matrix

$$\tilde{A} = \frac{1}{5} \begin{bmatrix} 3 & 4 \\ -4 & 3 \end{bmatrix}$$

is an orthogonal matrix.

Exercise 11 Let X_1, X_2, X_3 be the matrices in Exercise 3, and

$$C = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 0 & \sqrt{2} \\ 1 & 1 & 0 \end{bmatrix}.$$

Apply Theorem 4 and the solution of Exercise 3 to compute the singular values of the matrices X_1C, X_2C, and X_3C.

Hint: Consider the orthogonal matrix $\frac{1}{\sqrt{2}}C$ in the application of Theorem 4.

Exercise 12 Apply Theorem 5 to determine the singular values of the following square matrices simply by computing their eigenvalues. Justify the validity of this approach by applying Theorem 5.

(a) $A_1 = \begin{bmatrix} 1 & i \\ -i & 0 \end{bmatrix}$.

(b) $A_2 = \begin{bmatrix} 0 & 1+i \\ 1-i & 0 \end{bmatrix}$.

(c) $A_3 = \begin{bmatrix} -1 & 2 \\ 2 & 1 \end{bmatrix}$.

3.2 Matrix Norms and Low-Rank Matrix Approximation

When a matrix $B \in \mathbb{C}^{m,n}$ is considered as a linear transformation from $\mathbb{V} = \mathbb{C}^n$ to $\mathbb{W} = \mathbb{C}^m$, the norm of B is the operator norm $\||B|\|_{\mathbb{V} \to \mathbb{W}}$ (see (2.2.4) on p.180). Of course this definition depends on the choice of norms for the spaces $\mathbb{V} = \mathbb{C}^n$ and $\mathbb{W} = \mathbb{C}^m$. If there is no indication of such choice, then it is a common practice to assume that the Euclidean (that is, ℓ_2^m and ℓ_2^n) norms are used. More generally, recall from Sect. 1.5 of Chap. 1, that for any p with $1 \le p \le \infty$, the normed (linear) space ℓ_p^n is the vector space $\mathbb{C}^n$ over $\mathbb{C}$ with norm $\| \cdot \|_p$, defined by

(a) for $1 \leq p < \infty :$ $\quad \|\mathbf{x}\|_p = \left(\sum_{j=1}^n |x_j|^p \right)^{\frac{1}{p}};$

(b) for $p = \infty :$ $\quad \|\mathbf{x}\|_\infty = \max\{|x_j| : 1 \leq j \leq n\},$

where $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{C}^n$. Of course for the real-valued setting, $\mathbb{C}^n$ is replaced by $\mathbb{R}^n$ and $\mathbb{C}$ by $\mathbb{R}$.

For simplicity, we will replace the notation $\|\|B\|\|_{\ell_p^n \to \ell_p^m}$, in (2.2.4) and (2.2.5) on p.80, by $\|\|B\|\|_p$ for matrix transformations $B \in \mathbb{C}^{m,n}$, as follows.

Definition 1 $\boxed{\textbf{Operator norm of matrices}}$ *Let $B \in \mathbb{C}^{m,n}$. For $1 \leq p \leq \infty$, the ℓ_p-operator norm of B is defined by*

$$\|\|B\|\|_p = \max \left\{ \|B\mathbf{x}\|_p : \mathbf{x} \in \mathbb{C}^n, \|\mathbf{x}\|_p = 1 \right\} \tag{3.2.1}$$
$$= \max \left\{ \frac{\|B\mathbf{x}\|_p}{\|\mathbf{x}\|_p} : \mathbf{x} \in \mathbb{C}^n, \mathbf{x} \neq \mathbf{0} \right\}.$$

Similarly, for $B \in \mathbb{R}^{m,n}$, $\|\|B\|\|_p$ is defined as in (3.2.1) with $\mathbb{C}^n$ replaced by $\mathbb{R}^n$.

To be precise, we point out that the ℓ_p-norm in (3.2.1) for $B\mathbf{x}$ and $\mathbf{x}$ are somewhat different, since it stands for the ℓ_p^m-norm of $B\mathbf{x} \in \mathbb{C}^m$ and the ℓ_p^n-norm of $\mathbf{x} \in \mathbb{C}^n$. Again, for the real-valued setting, we simply replace $\mathbb{C}^n$ by $\mathbb{R}^n$. From the definition (3.2.1), it is clear that

$$\|B\mathbf{x}\|_p \leq \|\|B\|\|_p \|\mathbf{x}\|_p, \text{ for all } \mathbf{x} \in \mathbb{C}^n, \tag{3.2.2}$$

and that $\|\|B\|\|_p \leq C$, for any constant $C > 0$ for which

$$\|B\mathbf{x}\|_p \leq C\|\mathbf{x}\|_p, \ \mathbf{x} \in \mathbb{C}^n. \tag{3.2.3}$$

In other words, $\|\|B\|\|_p$ in (3.2.2) is the smallest constant C in (3.2.3).

In the following, we focus our attention on $\|\|B\|\|_p$, only for $p = 1, 2, \infty$. For the Euclidean space, we have the following result.

Theorem 1 $\boxed{\|\|\mathbf{B}\|\|_2 \textbf{ norm } = \textbf{largest singular value of B}}$ *For any $B \in \mathbb{C}^{m,n}$, its operator norm $\|\|B\|\|_2$, as defined by (3.2.1) with $p = 2$, is precisely the largest singular value of B; that is,*

$$\|\|B\|\|_2 = \sigma_1, \tag{3.2.4}$$

where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_n = 0$ are the singular values of B, as introduced in (3.1.3) on p.117.

Proof To prove (3.2.4), consider the spectral decomposition $B^*B = V\Lambda V^*$ of B^*B, with unitary matrix V and $\Lambda = \text{diag}\{\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2\}$, where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$. Then for any $\mathbf{x} \in \mathbb{C}^n$, we have

$$\langle B\mathbf{x}, B\mathbf{x} \rangle = \langle \mathbf{x}, B^*B\mathbf{x} \rangle = \langle \mathbf{x}, V\Lambda V^*\mathbf{x} \rangle = \langle V^*\mathbf{x}, \Lambda V^*\mathbf{x} \rangle.$$

Hence, applying the Cauchy-Schwarz inequality, we obtain

$$\|B\mathbf{x}\|_2^2 = \langle B\mathbf{x}, B\mathbf{x}\rangle = \langle V^*\mathbf{x}, \Lambda V^*\mathbf{x}\rangle \le \|V^*\mathbf{x}\|_2 \, \|\Lambda V^*\mathbf{x}\|_2$$
$$\le \|V^*\mathbf{x}\|_2 \, \sigma_1^2 \|V^*\mathbf{x}\|_2 = \sigma_1^2 \|\mathbf{x}\|_2^2,$$

where the second inequality follows from the fact that $\|\Lambda\mathbf{y}\|_2 \le \sigma_1^2\|\mathbf{y}\|_2$ for any $\mathbf{y} \in \mathbb{C}^n$ (see Exercise 12) and the last equality follows from the length invariant property of unitary transformations (see Theorem 2 on p.107). This yields $\|B\mathbf{x}\|_2^2 \le \sigma_1^2\|\mathbf{x}\|_2^2$ for any $\mathbf{x} \in \mathbb{C}^n$; and hence, it follows from (3.2.3) that $\||B\||_2 \le \sigma_1$.

On the other hand, by a suitable choice of the first column of V, namely: $\mathbf{x}_0 = V\mathbf{e}_1$, where $\mathbf{e}_1 = [1, 0, \ldots, 0]^T$, we have $\|\mathbf{x}_0\|_2 = 1$ and $V^*\mathbf{x}_0 = \mathbf{e}_1$, so that

$$\|B\mathbf{x}_0\|_2^2 = \langle V^*\mathbf{x}_0, \Lambda V^*\mathbf{x}_0\rangle = \langle \mathbf{e}_1, \Lambda\mathbf{e}_1\rangle = \sigma_1^2 = \sigma_1^2 \, \|\mathbf{x}_0\|_2^2,$$

which yields $\||B\||_2 \ge \sigma_1$, in view of (3.2.2). Therefore, by combining the two inequalities, we have proved (3.2.4).   ∎

In view of Theorem 1, it is natural to introduce the following definition.

Definition 2 ⎢**Spectral norm**⎥ *The operator norm $\||B\||_2$ of a matrix $B \in \mathbb{C}^{m,n}$ is called the spectral norm of B.*

For Hermitian matrices, the spectral norm is the largest eigenvalue (see Theorem 5 on p.130), which is characterized by the maximum of the Rayleigh quotients, as already pointed out in Remark 4 (see (2.3.19)) on p.101 in Sect. 2.2 of Chap. 2. This fact is a consequence of Theorem 1 and the following result.

Theorem 2 *If B is a Hermitian matrix, then*

$$\||B\||_2 = \max_{\|\mathbf{x}\|_2=1} |\langle B\mathbf{x}, \mathbf{x}\rangle|.$$

Proof To prove Theorem 2 for square matrices $B \in \mathbb{C}^{m,m}$, we first apply the Cauchy-Schwarz inequality to obtain

$$|\langle B\mathbf{x}, \mathbf{x}\rangle| \le \|B\mathbf{x}\| \, \|\mathbf{x}\| \le \||B\||_2 \, \|\mathbf{x}\|^2 \||B\||_2,$$

which holds for any $\mathbf{x} \in \mathbb{C}^m$ with $\|\mathbf{x}\| = 1$, so that

$$\max_{\|\mathbf{x}\|_2=1} |\langle B\mathbf{x}, \mathbf{x}\rangle| \le \||B\||_2.$$

To derive that $\||B\||_2$ is also a lower bound, we apply the spectral decomposition

$$B = U\Lambda U^*$$

of the Hermitian matrix B, with unitary matrix U and $\Lambda = \text{diag}\{\lambda_1, \ldots, \lambda_m\}$, where $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_m|$. Then since U is unitary, we have, for all $\mathbf{x} \in \mathbb{C}^m$,

$$
\begin{aligned}
\|B\mathbf{x}\|_2 &= \|U\Lambda U^*\mathbf{x}\|_2 = \|\Lambda U^*\mathbf{x}\|_2 \\
&\leq |\lambda_1| \, \|U^*\mathbf{x}\|_2 = |\lambda_1| \, \|U^*\mathbf{x}\|_2 = |\lambda_1| \, \|\mathbf{x}\|.
\end{aligned}
$$

Thus, it follows from the definition of the operator norm that

$$
\||B\||_2 \leq |\lambda_1|.
$$

On the other hand, if $\mathbf{v}_1$ is a unit eigenvector associated with λ_1, then

$$
|\langle B\mathbf{v}_1, \mathbf{v}_1\rangle| = |\langle \lambda_1\mathbf{v}_1, \mathbf{v}_1\rangle| = |\lambda_1| \, \|\mathbf{v}_1\|^2 = |\lambda_1|,
$$

which implies that

$$
|\lambda_1| \leq \max_{\|\mathbf{x}\|_2=1} |\langle B\mathbf{x}, \mathbf{x}\rangle|.
$$

The above two derivations together yield

$$
\||B\||_2 \leq |\lambda_1| \leq \max_{\|\mathbf{x}\|_2=1} |\langle B\mathbf{x}, \mathbf{x}\rangle|.
$$

Hence, $\||B\||_2$ is also a lower bound, completing the proof of the theorem. ∎

We remark that the above theorem can also be proved by applying Theorem 1 and the fact that $|\lambda_1| = \sigma_1$, which is assured by Theorem 5 on p.129 for normal matrices.

The **spectral radius** $\rho(B)$ of a square matrix B is defined as the largest eigenvalue (in modulus) of B:

$$
\rho(B) := |\lambda_1|.
$$

Theorem 3 *For a square matrix B,*

$$
\rho(B) = \lim_{k\to\infty} \||B^k\||^{\frac{1}{k}}, \tag{3.2.5}
$$

for any operator norm $\|| \cdot \||$.

The proof of (3.2.5) is beyond the scope of this book. Here, we only give the proof for the special case that B is diagonalizable; that is, there is a nonsingular matrix U such that

$$
B = U^{-1}DU,
$$

where $D = \text{diag}\{\lambda_1, \ldots, \lambda_n\}$ with λ_j being eigenvalues of B. Suppose $|\lambda_1| \geq \cdots \geq |\lambda_n|$. From

$$
B^k = U^{-1}D^kU,
$$

and

$$c_1 \, \||D^k|\|| \leq \||U^{-1}D^k U|\|| \leq c_2 \, \||D^k|\||,$$

where c_1, c_2 are some positive numbers, we have

$$\sqrt[k]{c_1} \, \||D^k|\||^{\frac{1}{k}} \leq \||B^k|\||^{\frac{1}{k}} \leq \sqrt[k]{c_2} \, \||D^k|\||^{\frac{1}{k}}.$$

From this and the facts $\||D^k|\||^{\frac{1}{k}} = |\lambda_1|$, $\lim_{k\to\infty} \sqrt[k]{c_1} = \lim_{k\to\infty} \sqrt[k]{c_2} = 1$, we conclude that (3.2.5) holds. ∎

Next we will show that the operator norm $\||B|\||_p$, for $p = 1$ and $p = \infty$, can be computed in terms of the ℓ_1-norms of the columns and of the rows, of the matrix B, respectively.

Theorem 4 *For any matrix $B = [b_{jk}] \in \mathbb{C}^{m,n}$,*

$$\||B|\||_1 = \max_{1\leq k\leq n} \left\{ \sum_{j=1}^{m} |b_{jk}| \right\}; \tag{3.2.6}$$

$$\||B|\||_\infty = \max_{1\leq j\leq m} \left\{ \sum_{k=1}^{n} |b_{jk}| \right\}. \tag{3.2.7}$$

For the sake of easier reading, we first demonstrate the computation of $\||B|\||_p$, for $p = 1, 2, \infty$, with the following example, and delay the derivations of (3.2.6) and (3.2.7) to the end of this section.

Example 1 Compute the operator norms $\||B|\||_2$, $\||B|\||_1$ and $\||B|\||_\infty$ of the matrix

$$B = \begin{bmatrix} 0 & 1 & 1 \\ 3 & -1 & 1 \\ 3 & 1 & -1 \end{bmatrix}.$$

Solution By Theorem 1, to determine the spectral norm $\||B|\||_2$, we may compute the largest singular value σ_1 of B. For this purpose, first compute $A = BB^*$, namely:

$$A = BB^T = \begin{bmatrix} 0 & 1 & 1 \\ 3 & -1 & 1 \\ 3 & 1 & -1 \end{bmatrix} \begin{bmatrix} 0 & 3 & 3 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 11 & 7 \\ 0 & 7 & 11 \end{bmatrix}.$$

The eigenvalues of A can be computed from the determinant of $\lambda I_3 - A$, which is

$$(\lambda - 2) \begin{vmatrix} \lambda - 11 & -7 \\ -7 & \lambda - 11 \end{vmatrix} = (\lambda - 2)(\lambda^2 - 22\lambda + 72)$$

$$= (\lambda - 2)(\lambda - 4)(\lambda - 18),$$

with eigenvalues of A given by the roots

$$\lambda_1 = 18, \ \lambda_2 = 4, \lambda_3 = 2.$$

Then the singular values of B are the positive square-roots of the eigenvalues of A, namely:

$$\sigma_1 = \sqrt{18} = 3\sqrt{2}, \quad \sigma_2 = \sqrt{4} = 2, \quad \sigma_3 = \sqrt{2}.$$

Hence, the largest singular value is $\sigma_1 = 3\sqrt{2}$, so that

$$\|\|B\|\|_2 = \sigma_1 = 3\sqrt{2}$$

by Theorem 1.

To compute the operator norm $\|\|B\|\|_1$, we simply compute the ℓ_1-norm of each column of B, which is the sum of each column of the matrix

$$\widetilde{B} = [|b_{jk}|] = \begin{bmatrix} 0 & 1 & 1 \\ 3 & 1 & 1 \\ 3 & 1 & 1 \end{bmatrix},$$

obtained from B by taking the absolute value of its entries. The result is the set of three numbers 6, 3, 3. By (3.2.6) of Theorem 4, we have

$$\|\|B\|\|_1 = \max\{6, 3, 3\} = 6.$$

Similarly, to find the operator norm $\|\|B\|\|_\infty$, we compute the sum of each row of the matrix $\widetilde{B}$ associated with B. The result is the set of three numbers 2, 5, 5. Hence by (3.2.7) of Theorem 4, we have

$$\|\|B\|\|_\infty = \max\{2, 5, 5\} = 5.$$

<div align="right">■</div>

Next we introduce the notion of Frobenius norm (or Hilbert-Schmidt norm). In this case, a matrix $B = [b_{jk}] \in \mathbb{C}^{m,n}$ is considered as an $mn \times 1$-vector $\mathbf{b} \in \mathbb{C}^{mn,1}$, in that all the entries of B are arranged as a finite sequence, such as

$$\mathbf{b} = (b_{11}, \ldots, b_{m1}, b_{12}, \ldots, b_{m2}, \ldots, b_{1n}, \ldots, b_{mn}).$$

Then the Frobenius norm of B is defined by the Euclidean norm of the vector $\mathbf{b}$, as follows.

Definition 3 | **Frobenius norm** | *The Frobenius norm* (also called Hilbert-Schmidt norm) of an $m \times n$ matrix $B = [b_{jk}]$ is defined by

$$\|B\|_F = \Big(\sum_{j=1}^{m}\sum_{k=1}^{n}|b_{jk}|^2\Big)^{1/2}.$$

Of course the definition of the Frobenius norm can be generalized from the Euclidean ℓ_2-norm to ℓ_p, for any $1 \le p \le \infty$. It can even be generalized to $0 < p < 1$, but without taking the power of $1/p$, for the validity of the triangle inequality, as discussed in Sect. 1.2 of the first chapter. But we are only interested in the case $p = 2$, as in the above definition of the Frobenius norm, since it occurs to be the only useful consideration for the study of data dimensionality reduction, to be studied in the final section of this chapter.

Theorem 5 $\boxed{\textbf{Frobenius norm} = \ell_2\textbf{-norm of singular values}}$ *Let $B \in \mathbb{C}^{m,n}$ with* rank$(B) = r$. *Then*

$$\|B\|_F = \Big(\sum_{j=1}^{r}\sigma_j^2\Big)^{1/2}, \tag{3.2.8}$$

where $\sigma_1, \dots, \sigma_r$ are the singular values of B.

Proof To derive (3.2.8), let $A = BB^*$ and observe that the Frobenius norm $\|B\|_F$ of B agrees with the trace of $A = [a_{jk}]$, $j = 1, \dots, n$, introduced in Definition 2 on p.93. This fact is a simple consequence of the definition of the trace, namely:

$$\begin{aligned}
\mathrm{Tr}(A) &= \sum_{k=1}^{n} a_{k,k} = \sum_{k=1}^{n}\Big(\sum_{\ell=1}^{m} b_{k,\ell}\overline{b_{k,\ell}}\Big) \\
&= \sum_{k=1}^{n}\sum_{\ell=1}^{m}|b_{\ell,k}|^2 = \|B\|_F^2.
\end{aligned}$$

On the other hand, recall from (2.3.5) of Theorem 1 on p.90 that $\mathrm{Tr}(A)$ is the sum of the eigenvalues of A. By the definition of the singular values of B, these eigenvalues

$$\lambda_1 = \sigma_1^2, \ \lambda_2 = \sigma_2^2, \ \dots, \ \lambda_n = \sigma_n^2$$

of A are squares of the singular values of the matrix B. Hence, we may conclude that

$$\|B\|_F = \big(\mathrm{Tr}(A)\big)^{1/2} = \Big(\sum_{j=1}^{n}\sigma_j^2\Big)^{1/2} = \Big(\sum_{j=1}^{r}\sigma_j^2\Big)^{1/2},$$

since $\sigma_{r+1} = \dots = \sigma_m = 0$. ∎

Example 2 Let B be the matrix considered in Example 1. Compute the Frobenius norm $\|B\|_F$ of B.

Solution The Frobenius norm $\|B\|_F$ can be computed directly by applying the definition, namely:

$$\|B\|_F^2 = \sum_{j=1}^{3} \sum_{k=1}^{3} |b_{jk}|^2 = 0^2 + 3^2 + 3^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2 = 24,$$

so that $\|B\|_F = \sqrt{24} = 2\sqrt{6}$.

Since the singular values of B have already been obtained in Example 1, we may also apply (3.2.8) of Theorem 5 to compute $\|B\|_F$, namely:

$$\|B\|_F = \left(\sum_{j=1}^{3} \sigma_j^2 \right)^{\frac{1}{2}} = (18 + 4 + 2)^{\frac{1}{2}} = 2\sqrt{6},$$

which agrees with the above answer obtained by using the definition. ∎

Applying Theorem 5 to generalize the Frobenius norm, from the ℓ_2-norm to the general ℓ_p-norm, of the sequence of singular values B, as opposed to the sequence of all of the entries of the matrix B, we introduce the following notion of Schatten norm.

Definition 4 ⎸**Schatten p-norm**⎹ *Let $B \in \mathbb{C}^{m,n}$ with* $\mathrm{rank}(B) = r$. *For* $1 \le p \le \infty$, *the Schatten p-norm of B is defined as*

$$\|B\|_{*,p} = \left(\sum_{j=1}^{r} \sigma_j^p \right)^{1/p},$$

where $\sigma_1, \ldots, \sigma_r$ are the singular values of B.

Remark 5 The Schatten p-norm $\|B\|_{*,p}$ not only generalizes the Frobenius norm, but also the spectral norm as well. Indeed, $\|B\|_{*,2} = \|B\|_F$ and $\|B\|_{*,\infty} = \|\|B\|\|_2$, since $\|B\|_{*,\infty} = \sigma_1$ and by Theorem 1, $\|\|B\|\|_2 = \sigma_1$ also. In addition, for $p = 1$, the Schatten 1-norm $\|B\|_{*,1} = \sum_{j=1}^{r} \sigma_j$ is called the **nuclear norm** (also called **trace norm** and **Ky Fan norm** in the literature). Since the nuclear norm is useful for low-rank and sparse matrix decomposition, a subject to be studied in a forthcoming publication of this book series, we use the abbreviated notation

$$\|B\|_* = \|B\|_{*,1} = \sum_{j=1}^{r} \sigma_j \tag{3.2.9}$$

for simplicity. ∎

Remark 6 When $\mathbb{C}^{m,n}$ is considered as a vector space over the scalar field $\mathbb{C}$, both the operator norm $\|\|B\|\|_p$ and Schatten norm $\|B\|_{*,p}$, for any $1 \le p \le \infty$, provide

norm measurements of $B \in \mathbb{V}$, meaning that they satisfy the norm properties (a)–(c) of Definition 2 on p.55. While verification of these norm properties for the operator norm is fairly simple for all $1 \leq p \leq \infty$ (see Exercise 14–15), proof of the triangle inequality:

$$\|A + B\|_{*,p} \leq \|A\|_{*,p} + \|B\|_{*,p} \quad \text{where } A, B \in \mathbb{C}^{m,n},$$

for the Schatten p-norm, with p different from $1, 2$ and ∞, is somewhat tedious and not provided in this book. We also mention that the operator norm satisfies the sub-multiplicative condition, meaning that

$$\|\|AB\|\|_p \leq \|\|A\|\|_p \, \|\|B\|\|_p \quad \text{for all } A \in \mathbb{C}^{\ell,m} \text{ and } B \in \mathbb{C}^{m,n}$$

(see Exercise 16). Of course, all of the above statements remain valid for the real-valued setting, if we replace $\mathbb{C}^{m,n}$ by $\mathbb{R}^{m,n}$, $\mathbb{C}^{\ell,m}$ by $\mathbb{R}^{\ell,m}$, and $\mathbb{C}$ by $\mathbb{R}$. ∎

In the following, we establish the unitary invariance property of the Schatten p-norm.

Theorem 6 | **Unitary invariance of Schatten p-norm** | *For any p, with $1 \leq p \leq \infty$, and any $m \times n$ matrix B, the Schatten p-norm of arbitrary unitary transformations of B remains the same as the Schatten p-norm of B. More precisely,*

$$\|WBR\|_{*,p} = \|B\|_{*,p}$$

for all unitary matrices W and R of dimension $m \times m$ and $n \times n$, respectively.

The proof of Theorem 6 is accomplished simply by applying the unitary invariance property of singular values in Theorem 4 on p.126. ∎

Since $\|B\|_{*,2} = \|B\|_F$ (see Remark 1), Theorem 6 is valid for the Frobenius norm, namely: for any $m \times n$ matrix B,

$$\|WBR\|_F = \|B\|_F, \tag{3.2.10}$$

where W and R are arbitrary unitary matrices.

The Frobenius norm will be used to assess the exact error in the approximation of matrices B with rank r, by matrices C with rank $\leq d$, for any desired $d < r$. Approximation by lower-rank matrices is the first step to the understanding of data dimensionality reduction, a topic of applications to be discussed in the next section. The basic tool for this study is the following decomposition formula of any matrix $B \in \mathbb{C}^{m,n}$ with rank $= r$, obtained by applying singular value decomposition (SVD), namely:

$$B = U_1 \Sigma_r V_1^* = U S V^* = \sum_{j=1}^{r} \sigma_j \mathbf{u}_j \mathbf{v}_j^*, \tag{3.2.11}$$

where U, V are unitary matrices of dimensions m, n respectively, and $U_1 = [\mathbf{u}_1, \ldots, \mathbf{u}_r]$, $V_1 = [\mathbf{v}_1, \ldots, \mathbf{v}_r]$ are obtained from U, V by keeping only the first r columns, where $\sigma_1 \geq \cdots \geq \sigma_r > 0$ are the non-zero singular values of B (with multiplicities being listed), and $\mathbf{u}_j^*, \mathbf{v}_j^*$ denote the complex conjugate of the transpose of the jth columns $\mathbf{u}_j, \mathbf{v}_j$ of U_1, V_1 in (3.1.15) on p.120 or U, V in (3.1.19) on p.121, respectively. To derive (3.2.11), we simply apply the reduced SVD of B in (3.1.15) on p.120, re-formulate the two matrix components $U_1 \Sigma_r$ and V_1^*, and finally multiply the corresponding column and row sub-blocks, as follows:

$$B = U_1 \Sigma_r V_1^* = [\sigma_1 \mathbf{u}_1 \ \cdots \ \sigma_r \mathbf{u}_r] \begin{bmatrix} \mathbf{v}_1^* \\ \vdots \\ \mathbf{v}_r^* \end{bmatrix} = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^* + \cdots + \sigma_r \mathbf{u}_r \mathbf{v}_r^*.$$

For each $j = 1, \ldots, r$, observe that $\mathbf{u}_j \mathbf{v}_j^*$ is an $m \times n$ matrix with rank $= 1$. Such matrices are called rank-1 matrices. We remark that an $m \times n$ matrix B is a rank-1 matrix, if and only if $B = \mathbf{v}\mathbf{w}^T$, where $\mathbf{v}$ and $\mathbf{w}$ are m-dimensional and n-dimensional column vectors, respectively (see Exercises 6–7), and that the rank of the sum of r rank-1 $m \times n$ matrices does not exceed r (see Exercises 8–10).

Theorem 7 $\boxed{\textbf{Lower-rank matrix approximation}}$ *Let B be any $m \times n$ matrix of complex* (or real) *numbers with* $\text{rank}(B) = r$ *and with singular values* $\sigma_1 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = 0$. *Then for any integer d, with $1 \leq d \leq r$, the dth partial sum*

$$B(d) = \sum_{j=1}^{d} \sigma_j \mathbf{u}_j \mathbf{v}_j^*, \tag{3.2.12}$$

of the rank-1 matrix series representation (3.2.11) *of B, provides the best approximation of B by all matrices of rank $\leq d$ under the Frobenius norm, with precise error given by $\sigma_{d+1}^2 + \cdots + \sigma_r^2$; that is,*

$$\|B - B(d)\|_F^2 = \sum_{j=d+1}^{r} \sigma_j^2, \tag{3.2.13}$$

and

$$\|B - B(d)\|_F^2 \leq \|B - C\|_F^2 \tag{3.2.14}$$

for all $m \times n$ matrices C with rank $\leq d$. Furthermore, $B(d)$ in (3.2.12) is the unique best approximant of B under the Frobenius norm.

Proof Let S_d denote the matrix obtained from the matrix $S \in \mathbb{R}^{m,n}$, introduced in (3.1.6)–(3.1.7) on p.117, by replacing each of $\sigma_1, \ldots, \sigma_d$ with 0. Then we have

$$B - B(d) = U S_d V^*$$

(see Exercise 13). Hence, it follows from (3.2.10) that

$$\|B - B(d)\|_F^2 = \|U S_d V^*\|_F^2 = \|S_d\|_F^2 = \sum_{j=d+1}^{r} \sigma_j^2,$$

completing the derivation of (3.2.13).

To prove (3.2.14), assume that $C \in \mathbb{C}^{m,n}$ with rank $= k \le d$ provides the best approximation to B under the Frobenius norm $\|\cdot\|_F$, so that $\|B - C\|_F^2 \le \|B - B(d)\|_F^2$. Hence, it follows from (3.2.13) that

$$\|B - C\|_F^2 \le \sum_{j=d+1}^{r} \sigma_j^2. \tag{3.2.15}$$

Let U, V be the unitary matrices in the SVD of B in (3.2.11) and set $G = U^*CV$, so that G has the same rank k as C and that $C = UGV^*$. Hence, by applying (3.2.10), we have

$$\begin{aligned}
\|B - C\|_F &= \|USV^* - UGV^*\|_F \\
&= \|U(S - G)V^*\|_F = \|S - G\|_F.
\end{aligned}$$

Set $G = [g_{j,\ell}]$ and $S = [s_{j,\ell}]$, where in view of (3.1.6)–(3.1.7) on p.117, $s_{j,\ell} = 0$ for all j different from ℓ and $s_{j,j} = 0$ for $j > r$. Since the Frobenius norm $\|S - G\|_F$ is defined by the ℓ_2 sequence norm of the sequence consisting of all the entries $s_{j,\ell} - g_{j,\ell}$ of the matrix $S - G$, it should be clear (see Exercise 17) that for $\|S - G\|_F$ to be minimum among all $G \in \mathbb{C}^{m,n}$, this optimal G is given by

$$G = \begin{bmatrix} \Sigma_r' & 0 \\ 0 & 0 \end{bmatrix},$$

where $\Sigma_r' = \text{diag}\{g_1, g_2, \ldots, g_r\}$, with each $g_j \ge 0$, so that

$$\|S - G\|_F^2 = \sum_{j=1}^{r} |s_{j,j} - g_j|^2 = \sum_{j=1}^{r} |\sigma_j - g_j|^2.$$

Now, since the rank of G is $k \le d \le r$, only k of the r diagonal entries $g_1, g_2, \ldots, g_r$ of Σ_r' are non-zero, and the minimum of the above sum is achieved only when these k non-zero entries match the largest k values of $\sigma_1, \ldots, \sigma_r$ (see Exercise 18). In other words, we have $g_1 = \sigma_1, \ldots, g_k = \sigma_k$, and $g_j = 0$ for $j > k$, and that

$$\|B - C\|_F^2 = \|S - G\|_F^2 = \sum_{j=k+1}^{r} \sigma_j^2 \qquad (3.2.16)$$

(see Exercise 19). Hence, by combining (3.2.15) and (3.2.16), we may conclude that $k = d$,

$$\|B - C\|_F^2 = \sum_{j=d+1}^{r} \sigma_j^2,$$

and that $C = UGV^*$, with

$$G = \begin{bmatrix} \Sigma_r' & 0 \\ 0 & 0 \end{bmatrix},$$

where $\Sigma_r' = \mathrm{diag}\{\sigma_1, \ldots, \sigma_d\}$. This implies that $C = B(d)$ in (3.2.12), completing the proof of the theorem. ∎

We conclude this section by providing the following proof of Theorem 4.

Proof of Theorem 4 To derive the formula (3.2.6), let c_0 be defined by

$$c_0 := \sum_{j=1}^{m} |b_{jk_0}|,$$

and assume that this column-norm is the largest among all the column-norms $\sum_{j=1}^{m} |b_{jk}|$, for $1 \le k \le m$. Then for any $\mathbf{x} \in \mathbb{C}^n$, we have

$$\|B\mathbf{x}\|_1 = \sum_{j=1}^{m} \left| \sum_{k=1}^{n} b_{jk} x_k \right| \le \sum_{j=1}^{m} \sum_{k=1}^{n} |b_{jk}| \, |x_k|$$
$$= \sum_{k=1}^{n} \sum_{j=1}^{m} |b_{jk}| \, |x_k| \le \sum_{k=1}^{n} c_0 \, |x_k| = c_0 \|\mathbf{x}\|_1.$$

Since $\|\|B\|\|_1$ is the minimum among all upper bounds, we have

$$\|\|B\|\|_1 \le c_0.$$

On the other hand, by selecting the coordinate unit vector $\mathbf{x}_0 = \mathbf{e}_{k_0} = [0, \ldots, 0, 1, 0, \ldots, 0]^T \in \mathbb{R}^n$, we obtain

$$\|B\mathbf{x}_0\|_1 = \|[b_{1k_0}, b_{2k_0}, \ldots, b_{mk_0}]^T\|_1 = \sum_{j=1}^{m} |b_{jk_0}| = c_0 \|\mathbf{x}_0\|_1.$$

Thus, it follows from the definition of the operator norm $||| \cdot |||_1$ that $|||B|||_1 \geq c_0$. Since the value c_0, defined above, provides both an upper bound and a lower bound of $|||B|||_1$, we have

$$|||B|||_1 = c_0 = \max_{1 \leq k \leq n} \left\{ \sum_{j=1}^{m} |b_{jk}| \right\}.$$

To derive the formula (3.2.7) for $|||B|||_\infty$, let d_0 be defined by

$$d_0 := \sum_{k=1}^{n} |b_{j_0 k}|$$

and assume that this row-norm is the largest among all the row-norms $\sum_{k=1}^{n} |b_{jk}|$, for $1 \leq j \leq m$. Then for any $\mathbf{x} \in \mathbb{C}^n$, we have

$$\|B\mathbf{x}\|_\infty = \max_{1 \leq j \leq m} \left\{ \left| \sum_{k=1}^{n} b_{jk} x_k \right| \right\} \leq \max_{1 \leq j \leq m} \left\{ \sum_{k=1}^{n} |b_{jk}| \, |x_k| \right\}$$

$$\leq \max_{1 \leq j \leq m} \left\{ \sum_{k=1}^{n} |b_{jk}| \, \|\mathbf{x}\|_\infty \right\} = d_0 \, \|\mathbf{x}\|_\infty.$$

Since $|||B|||_\infty$ is the minimum among all upper bounds, we have

$$|||B|||_\infty \leq d_0.$$

Next, for any complex number z, we introduce the notation

$$\text{conj}(z) = \begin{cases} \bar{z}/|z|, & \text{if } z \neq 0 \\ 0, & \text{if } z = 0, \end{cases}$$

so that $z \, \text{conj}(z) = |z|$.

Then by selecting the vector $\mathbf{x}_0 = [\text{conj}(b_{j_0 1}), \text{conj}(b_{j_0 2}), \ldots, \text{conj}(b_{j_0 n})]^T$, we obtain

$$\|B\mathbf{x}_0\|_\infty = \max_{1 \leq j \leq m} \left\{ \left| \sum_{k=1}^{n} b_{jk} \, \text{conj}(b_{j_0 k}) \right| \right\}$$

$$\geq \left| \sum_{k=1}^{n} b_{j_0 k} \, \text{conj}(b_{j_0 k}) \right| = \sum_{k=1}^{n} |b_{j_0 k}| = d_0 \|\mathbf{x}_0\|_\infty,$$

since $\|\mathbf{x}_0\|_\infty = 1$. Thus, it follows from the definition of operator norm that $|||B|||_\infty \geq d_0$. Since the value d_0 defined above is both an upper bound and a lower bound, we may conclude that

$$\||B\||_\infty = d_0 = \max_{1 \le j \le m} \left\{ \sum_{k=1}^{n} |b_{jk}| \right\}.$$

∎

Exercises

Exercise 1 Compute the operator norm $\||\cdot\||_2$ of each of the following matrices:

(a) $B = \begin{bmatrix} 2 & 1 & 1 \\ 1 & -1 & 1 \end{bmatrix}$,

(b) $C = \begin{bmatrix} 2 & 1 & 1 \\ 1 & -1 & 1 \\ 2 & -1 & -3 \end{bmatrix}$.

Exercise 2 Determine the spectral norms $\||\cdot\||_1$ and $\||\cdot\||_\infty$ of each of the matrices in Exercise 1.

Exercise 3 Find the Frobenius norm $\|\cdot\|_F$ of each of the matrices in Exercise 1.

Exercise 4 Determine the nuclear norm $\|\cdot\|_*$ of each of the matrices in Exercise 1.

Exercise 5 What is the spectral radius of the matrix C in Exercise 1?

Exercise 6 Let $\mathbf{x} \in \mathbb{C}^m$ and $\mathbf{y} \in \mathbb{C}^n$ be arbitrarily given non-zero column vectors. Show that the $m \times n$ matrix $B = \mathbf{x}\mathbf{y}^T$ is a rank-1 matrix; that is, $\text{rank}(B) = 1$.

Exercise 7 Let B be an $m \times n$ matrix of complex numbers. Show that if $\text{rank}(B) = 1$, then $B = \mathbf{x}\mathbf{y}^T$ for some vectors $\mathbf{x} \in \mathbb{C}^m$ and $\mathbf{y} \in \mathbb{C}^n$.

Exercise 8 Let B_1 and B_2 be rank-1 matrices. Show that $\text{rank}(B_1 + B_2) \le 2$.

Exercise 9 As an extension of Exercise 8, let B be a rank-1 matrix and $A = B + C$. Then show that $\text{rank}(A) \le \text{rank}(C) + 1$.

Exercise 10 As another extension of Exercise 8 (as well as Exercise 9), show that

$$\text{rank}\left(\sum_{j=1}^{r} \mathbf{x}_j \mathbf{y}_j^T \right) \le r$$

for all $\mathbf{x}_1, \ldots, \mathbf{x}_r \in \mathbb{C}^m$ and $\mathbf{y}_1, \ldots, \mathbf{y}_r \in \mathbb{C}^n$.

Exercise 11 Compute the rank of each of the following matrices.

(a) $A_1 = \mathbf{x}_1 \mathbf{y}_1^T + \mathbf{x}_2 \mathbf{y}_2^T,$

where

$$\mathbf{x}_1 = [1\ 0\ -1]^T, \quad \mathbf{x}_2 = [0\ 1\ 2]^T,$$
$$\mathbf{y}_1^T = [2\ 1], \quad \mathbf{y}_2^T = [1\ -1].$$

(b)
$$A_2 = A_1 + \mathbf{x}_3 \mathbf{y}_3^T,$$

where A_1 is given in (a), and

$$\mathbf{x}_3 = [\, 2 \; -1 \; 0 \,]^T, \quad \mathbf{y}_3^T = [\, 0 \; 1 \,]^T.$$

Exercise 12 Let $A = \text{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$, where $\lambda_j \in \mathbb{C}$ and $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|$. Show that

$$\|A\mathbf{x}\| \leq |\lambda_1| \, \|\mathbf{x}\|$$

for any $\mathbf{x} \in \mathbb{C}^n$, and that $\||A|\|_2 = |\lambda_1|$.

Exercise 13 Let S_d be the matrix obtained from the matrix S in (3.1.6)–(3.1.7) on p.117 by replacing each of $\sigma_1, \ldots, \sigma_d$ with 0. Show that $B - B(d) = U S_d V^*$.

Exercise 14 For each p with $1 \leq p \leq \infty$, show that the vector space of all matrices $B \in \mathbb{C}^{m,n}$ endowed with the operator norm $\||B|\|_p$ is a normed vector space, by verifying the norm properties (a)–(c) of Definition 2 on p.55.

Exercise 15 For each p with $1 \leq p \leq \infty$, verify that the Schatten p-norm $\|B\|_{*,p}$ of matrices $B \in \mathbb{C}^{m,n}$ satisfy the two properties (a) and (b) of Definition 2 on p.55.

Exercise 16 Show that the operator norm satisfies the sub-multiplicative condition:

$$\||AB|\|_p \leq \||A|\|_p \, \||B|\|_p,$$

for all matrices $A \in \mathbb{C}^{\ell,m}$ and $B \in \mathbb{C}^{m,n}$.

Exercise 17 Let $D \in \mathbb{C}^{n,n}$ be a diagonal matrix with rank $r > 0$ and $0 \leq k < r$. Suppose that C provides the best approximation of D from the collection of all matrices in $\mathbb{C}^{n,n}$ with rank $\leq k$ under the Frobenius norm; that is, $\|D - C\|_F \leq \|D - A\|_F$ for all matrices $A \in \mathbb{C}^{n,n}$ with rank $\leq k$. Show that C is also a diagonal matrix.

Exercise 18 As a continuation of Exercise 17, show that the entries of the diagonal matrix C, which best approximates D, are precisely the k entries of D with largest absolute values. Also show that all the diagonal entries of D are non-zero and that the rank of C is equal to k.

Exercise 19 As a continuation of Exercises 17 and 18, show that the approximant C of D is unique, and formulate the error of approximation $\|D - C\|_F$ in terms of the entries of the given matrix D.

3.3 Data Approximation

In this section, the mathematical development studied in the previous sections is applied to the problem on generalization of matrix inversion with applications

to solving arbitrary systems of linear equations and least-squares data estimation. Consider the matrix formulation

$$B\mathbf{x} = \mathbf{b}$$

of a given system of linear equations, with an arbitrary $m \times n$ coefficient matrix B, the (unknown) column $\mathbf{x}$, and any m-dimensional (known) column vector $\mathbf{b}$. Of course if B is a non-singular square matrix, then the solution of this system is simply $\mathbf{x} = B^{-1}\mathbf{b}$. But if $m > n$ (or $m < n$), the system is "over-determined" (or "under-determined") in the sense that there are more equations than unknowns (or more unknowns than equations), and the system could even be "inconsistent" in both cases. In the following, we provide the "optimal" solution of the (possibly inconsistent) system, by introducing the notion of "pseudo-inverse" $B^{\dagger}$ of B based on the SVD of the matrix B.

Definition 5 $\boxed{\text{Pseudo-inverse}}$ *Let B be an $m \times n$ matrix of real or complex numbers and $B = USV^*$ be its SVD, with unitary matrices U, V and S as given by (3.1.6) on p.117 with diagonal sub-block $\Sigma_r = \text{diag}\{\sigma_1, \ldots, \sigma_r\}$. Set $\Sigma_r^{-1} = \text{diag}\{\sigma_1^{-1}, \ldots, \sigma_r^{-1}\}$ and define the $n \times m$ matrix $\widetilde{S}$ by*

$$\widetilde{S} = \begin{bmatrix} \Sigma_r^{-1} & \vdots & O \\ \cdots & \cdots & \cdots \\ O & \vdots & O \end{bmatrix}_{n \times m}.$$

Then the $n \times m$ matrix

$$B^{\dagger} = V\widetilde{S}U^* \tag{3.3.1}$$

is called the pseudo-inverse of the given matrix B.

As in the above definition, the subscript $n \times m$ of the matrix of $\widetilde{S}$ is, and will be, used to indicate the matrix dimension. Observe that

$$BB^{\dagger} = (USV^*)(VS^+U^*) = U \begin{bmatrix} I_r & \vdots & O \\ \cdots & \cdots & \cdots \\ O & \vdots & O \end{bmatrix}_{m \times m} U^*$$

and

$$B^{\dagger}B = (VS^+U^*)(USV^*) = V \begin{bmatrix} I_r & \vdots & O \\ \cdots & \cdots & \cdots \\ O & \vdots & O \end{bmatrix}_{n \times n} V^*$$

are $m \times m$ and $n \times n$ square matrices, respectively, with $r \times r$ identity matrix sub-block I_r, where $r = \text{rank}(B)$. Hence, for non-singular square matrices, the pseudo-inverse

agrees with the inverse matrix. For this reason, the pseudo-inverse is also called generalized inverse.

To "solve" the following (possibly over-determined, under-determined, or even inconsistent) system of linear equations

$$B\mathbf{x} = \mathbf{b}, \tag{3.3.2}$$

(to be called a "linear system" for convenience), where B is an $m \times n$ coefficient matrix and $\mathbf{b}$ an m-dimensional (known) column vector, we may apply the pseudo-inverse $B^\dagger$ of B to $\mathbf{b}$ to formulate the vector:

$$\mathbf{x}^\diamond = B^\dagger \mathbf{b}, \tag{3.3.3}$$

and obtain the following result, where for a vector $\mathbf{y} \in \mathbb{C}^n$, its norm $\|\mathbf{y}\|$ denotes the Euclidean (or ℓ_2^n) norm of $\mathbf{y}$.

Theorem 8 $\boxed{\textbf{Pseudo-inverse provides optimal solution}}$ *For the linear system* (3.3.2) *with coefficient matrix* $B \in \mathbb{C}^{m,n}$ *and (known)* $\mathbf{b} \in \mathbb{C}^m$, *the vector* $\mathbf{x}^\diamond$ *defined in* (3.3.3) *has the following properties:*

(i) for all $\mathbf{x} \in \mathbb{C}^n$,
$$\|B\mathbf{x}^\diamond - \mathbf{b}\| \le \|B\mathbf{x} - \mathbf{b}\|;$$

(ii) the linear system (3.3.2), with unknown $\mathbf{x}$, has a solution if and only if the pseudo-inverse $B^\dagger$ of B satisfies the condition: $BB^\dagger \mathbf{b} = \mathbf{b}$, namely, $\mathbf{x} = \mathbf{x}^\diamond$ is a solution;

(iii) if (3.3.2) has a solution, then the general solution of (3.3.2) $\mathbf{x} \in \mathbb{C}^n$ is given by

$$\mathbf{x} = \mathbf{x}^\diamond + (I_n - B^\dagger B)\mathbf{w},$$

for all $\mathbf{w} \in \mathbb{C}^n$;

(iv) if (3.3.2) has a solution, then among all solutions, $\mathbf{x}^\diamond$ is the unique solution with the minimal Euclidean norm, namely:

$$\|\mathbf{x}^\diamond\| \le \|\mathbf{x}\|$$

for any solution $\mathbf{x}$ of (3.3.2); and

(v) if (3.3.2) has a unique solution, then $\text{rank}(B) = n$.

The above statements remain valid, when $\mathbb{C}^{m,n}, \mathbb{C}^n, \mathbb{C}^m$ are replaced by $\mathbb{R}^{m,n}, \mathbb{R}^n, \mathbb{R}^m$, respectively.

To prove the above theorem, we need the following properties of the pseudo-inverse.

Theorem 9 $\boxed{\textbf{Properties of pseudo-inverse}}$ *Let* $B \in \mathbb{C}^{n,m}$ *or* $B \in \mathbb{R}^{n,m}$. *Then*

(i) $(BB^\dagger)^* = BB^\dagger$;

(ii) $(B^\dagger B)^* = B^\dagger B$;

(iii) $BB^\dagger B = B$; *and*

(iv) $B^\dagger BB^\dagger = B^\dagger$.

Furthermore, $B^\dagger$ as defined by (3.3.1)*, is the only $n \times m$ matrix that satisfies the above conditions* (i)–(iv).

Proof Derivation of the properties (i)–(iv) of $B^\dagger$ is left as exercises (see Exercise 1). To show that $B^\dagger$ is unique, let $A \in \mathbb{C}^{m,n}$ satisfy (i)–(iv); that is,

(i) $(BA)^* = BA$;

(ii) $(AB)^* = AB$;

(iii) $BAB = B$; and

(iv) $ABA = A$.

In view of the definition $B^\dagger = V\widetilde{S}U^*$ of $B^\dagger$ in (3.3.1), we introduce four matrix sub-blocks $A_{11}, A_{12}, A_{21}, A_{22}$ of dimensions $r \times r, r \times (n - r), (m - r) \times r, (m - r) \times (n - r)$, respectively, defined by

$$V^*AU = \left[\begin{array}{ccc} A_{11} & \vdots & A_{12} \\ \cdots & \cdots & \cdots \\ A_{21} & \vdots & A_{22} \end{array}\right].$$

Then by the SVD formulation $B = USV^*$ of the given matrix B, it follows from the assumption $BAB = B$ in (iii) that

$$(U^*BV)(V^*AU)(U^*BV) = U^*(BAB)V = U^*BV.$$

Hence, by the definition (3.1.5) of S on p.117, we have

$$\begin{bmatrix} \Sigma_r & O \\ O & O \end{bmatrix}\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}\begin{bmatrix} \Sigma_r & O \\ O & O \end{bmatrix} = \begin{bmatrix} \Sigma_r & O \\ O & O \end{bmatrix},$$

which is equivalent to $\Sigma_r A_{11} \Sigma_r = \Sigma_r$. This yields $A_{11} = \Sigma_r^{-1}$. By applying the assumptions (i) and (ii) on A above, respectively, it can be shown that $A_{12} = O$ and $A_{21} = O$ (see Exercises 2 and 3). Finally, by applying these results along with the assumption $ABA = A$ in (iv), a similar derivation yields $A_{22} = O$ (see Exercise 4). Hence,

$$A = V\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}U^* = V\begin{bmatrix} \Sigma_r^{-1} & O \\ O & O \end{bmatrix}U^* = B^\dagger.$$

This completes the proof of the theorem. ∎

Proof of Theorem 1 To prove Theorem 1, write $B\mathbf{x}-\mathbf{b}=(B\mathbf{x}-B\mathbf{x}^\diamond)+(B\mathbf{x}^\diamond-\mathbf{b})$, and observe that the two vectors $B\mathbf{x} - B\mathbf{x}^\diamond$ and $B\mathbf{x}^\diamond - \mathbf{b}$ are orthogonal to each

other. The reason is that

$$(B\mathbf{x}^\diamond - \mathbf{b})^*(B\mathbf{x} - B\mathbf{x}^\diamond) = (BB^\dagger\mathbf{b} - \mathbf{b})^*B(\mathbf{x} - \mathbf{x}^\diamond)$$
$$= \mathbf{b}^*\Big((BB^\dagger)^* - I\Big)B(\mathbf{x} - \mathbf{x}^\diamond) = \mathbf{b}^*\Big(BB^\dagger B - B\Big)(\mathbf{x} - \mathbf{x}^\diamond) = 0,$$

where the last two equalities follow from (i) and (iii) of Theorem 2, respectively. Thus, it follows from the Pythagorean theorem that

$$\|B\mathbf{x} - \mathbf{b}\|^2 = \|B\mathbf{x} - B\mathbf{x}^\diamond\|^2 + \|B\mathbf{x}^\diamond - \mathbf{b}\|^2 \geq \|B\mathbf{x}^\diamond - \mathbf{b}\|^2,$$

establishing statement (i) in Theorem 1.

Statement (ii) follows immediately from statement (i), since if the system (3.3.2) has some solution, then $\mathbf{x}^\diamond$ in (3.3.1) is also a solution of (3.3.2).

To derive statement (iii), we first observe, in view of $BB^\dagger B = B$ in (iii) of Theorem 2, that for any $\mathbf{w} \in \mathbb{C}^n$, the vector $\mathbf{x} = \mathbf{x}^\diamond + (I_n - B^\dagger B)\mathbf{w}$ is a solution of (3.3.2), since (3.3.2) has a solution, namely, $\mathbf{x}^\diamond$. On the other hand, suppose that $\mathbf{x}$ is a solution. Then by setting $\mathbf{w} = \mathbf{x} - \mathbf{x}^\diamond$, we have $B\mathbf{w} = \mathbf{0}$, so that

$$\mathbf{x}^\diamond + (I_n - B^\dagger B)\mathbf{w} = \mathbf{x}^\diamond + \mathbf{w} - B^\dagger B\mathbf{w} = \mathbf{x}^\diamond + \mathbf{w} = \mathbf{x};$$

which establishes statement (iii).

To prove statement (iv), we apply (ii) and (iv) of Theorem 2 to show that in statement (iii) of the theorem, the vector $\mathbf{x}^\diamond$ is orthogonal to $(I_n - B^\dagger B)\mathbf{w}$, namely:

$$\Big((I_n - B^\dagger B)\mathbf{w}\Big)^*\mathbf{x}^\diamond = \mathbf{w}^*\Big(I_n - (B^\dagger B)^*\Big)B^\dagger\mathbf{b} = \mathbf{w}^*\Big(B^\dagger - B^\dagger BB^\dagger\Big)\mathbf{b} = \mathbf{0}.$$

Hence, since every solution $\mathbf{x}$ can be written as $\mathbf{x} = \mathbf{x}^\diamond + (I_n - B^\dagger B)\mathbf{w}$, we may apply the Pythagorean theorem to conclude that

$$\|\mathbf{x}\|^2 = \|\mathbf{x}^\diamond + (I_n - B^\dagger B)\mathbf{w}\|^2$$
$$= \|\mathbf{x}^\diamond\|^2 + \|(I_n - B^\dagger B)\mathbf{w}\|^2 \geq \|\mathbf{x}^\diamond\|^2.$$

Thus, $\mathbf{x}^\diamond$ is the unique solution of (3.3.2) with minimal norm.

Finally, to see that statement (v) holds, we simply observe that for the solution $\mathbf{x}^\diamond$ of (3.3.2) to be unique, the matrix $(I_n - B^\dagger B)$ in the general solution (in statement (iii)) must be the zero matrix; that is, $B^\dagger B = I_n$ or $B^\dagger$ is the right inverse of B, so that $\mathrm{rank}(B) = n$.   ∎

Example 3 As a continuation of Example 2 on p.118 in Sect. 3.1, compute the pseudo-inverse of the matrix

$$B = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & -1 \end{bmatrix}.$$

Solution Recall from Example 2 in Sect. 3.1 that the SVD of B is given by

$$B = USV^* = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \sqrt{2} & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}} \\ 0 & 1 & 0 \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \end{bmatrix}.$$

Hence, by the definition (3.3.1) of the pseudo-inverse $B^\dagger$ of B, we have

$$B^\dagger = V \widetilde{S} U^* = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ 0 & 1 & 0 \\ -\frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & \frac{1}{2} \\ 1 & 0 \\ 0 & -\frac{1}{2} \end{bmatrix}.$$

∎

Example 4 Write the system of linear equations

$$\begin{cases} x_2 = 3, \\ x_1 - x_3 = -1, \end{cases}$$

in matrix formulation $B\mathbf{x} = \mathbf{b}$ with $\mathbf{x} = [x_1, x_2, x_3]^T$ and $\mathbf{b} = [3, -1]^T$. Apply the result from Example 1 to obtain the solution $\mathbf{x}^\diamond = B^\dagger \mathbf{b}$ and verify that $||\mathbf{x}^\diamond|| \le ||\mathbf{x}||$ for all solutions $\mathbf{x}$ of the linear system.

Solution Since the coefficient matrix B is the matrix in Example 1, we may apply $B^\dagger$ computed above to obtain the solution

$$\mathbf{x}^\diamond = B^\dagger \mathbf{b} = \begin{bmatrix} 0 & \frac{1}{2} \\ 1 & 0 \\ 0 & -\frac{1}{2} \end{bmatrix} \begin{bmatrix} 3 \\ -1 \end{bmatrix} = \begin{bmatrix} -\frac{1}{2} \\ 3 \\ \frac{1}{2} \end{bmatrix}.$$

Furthermore, it is clear that the general solution of the system is

$$\mathbf{x} = [a - 1, 3, a]^T, \quad \text{for any real number } a,$$

so that

$$||\mathbf{x}||^2 = (a - 1)^2 + 3^2 + a^2 = a^2 - 2a + 1 + 9 + a^2$$

$$= 2(a^2 - a) + 10 = 2\left(a^2 - a + \tfrac{1}{4}\right) + 10 - \tfrac{2}{4}$$

$$= 2\left(a - \tfrac{1}{2}\right)^2 + \left(3^2 + \left(\tfrac{-1}{2}\right)^2 + \left(\tfrac{1}{2}\right)^2\right)$$

$$= 2\left(a - \tfrac{1}{2}\right)^2 + ||\mathbf{x}^\diamond||^2 \ge ||\mathbf{x}^\diamond||^2,$$

with $||\mathbf{x}|| = ||\mathbf{x}^\diamond||$ if and only if $a = \frac{1}{2}$, or $\mathbf{x} = \mathbf{x}^\diamond$.   ∎

Example 5 Consider the inconsistent system of linear equations

$$\begin{cases} x_2 = 1, \\ x_1 = -1, \\ -x_2 = 1, \end{cases}$$

with matrix formulation $B\mathbf{x} = \mathbf{b}$, where

$$B = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 0 & -1 \end{bmatrix} \quad \text{and } \mathbf{b} = \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}.$$

What would be a "reasonable" solution of the system?

Solution From Example 1 (see Example 2 in Sect. 3.1), since the coefficient matrix B is the transpose of the matrix B in Example 1, we have the SVD, $B = U S V^*$ with

$$U = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ 0 & 1 & 0 \\ -\frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \end{bmatrix},$$

$$V = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad S = \begin{bmatrix} \sqrt{2} & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}.$$

Hence, the pseudo-inverse $B^\dagger$ of B is given by

$$B^\dagger = V \tilde{S} U^* = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}} \\ 0 & 1 & 0 \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \end{bmatrix};$$

so that

$$\mathbf{x}^\diamond = B^\dagger \mathbf{b} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ -1 \\ \sqrt{2} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ -1 \end{bmatrix} = \begin{bmatrix} -1 \\ 0 \end{bmatrix}.$$

That is, $x_1 = -1$ and $x_2 = 0$ is the "reasonable" solution of the inconsistent system. Observe that the average of the inconsistency $x_2 = 1$ and $-x_2 = 1$ is the "reasonable" solution $x_2 = 0$.   ∎

Next, we apply Theorem 1 to study the problem of least-squares estimation.

Let $\mathbb{V}$ be an inner-product space over the scalar field $\mathbb{C}$ or $\mathbb{R}$ and $S_n = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ be a (possibly linearly dependent) set of vectors in $\mathbb{V}$ with $\mathbb{W} = \text{span } \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$. Since the cardinality n of the set S_n can be very large, to find a satisfactory representation of an arbitrarily given $\mathbf{v} \in \mathbb{V}$ from $\mathbb{W}$, it is often feasible to acquire only a subset of measurements $\langle \mathbf{v}, \mathbf{v}_\ell \rangle$ for $\ell \in \{n_1, \ldots, n_m\} \subset \{1, \ldots, n\}$. Let

$$\mathbf{b} = [b_1, \ldots, b_m]^T = [\langle \mathbf{v}, \mathbf{v}_{n_1} \rangle, \ldots, \langle \mathbf{v}, \mathbf{v}_{n_m} \rangle]^T \qquad (3.3.4)$$

be the data vector in $\mathbb{C}^m$ associated with $\mathbf{v}$ (where $m \leq n$). The least-squares estimation problem is to identify the "best" approximants

$$\mathbf{w} = \sum_{j=1}^{n} x_j \mathbf{v}_j \in \mathbb{W}$$

of the vector $\mathbf{v}$, based only on the measurement $\mathbf{b}$ in (3.3.4). Now, since $\langle \mathbf{w}, \mathbf{v}_{n_\ell} \rangle = \sum_{j=1}^{n} \langle \mathbf{v}_j, \mathbf{v}_{n_j} \rangle x_j$ is supposed to "match" the data component $\langle \mathbf{v}, \mathbf{v}_{n_\ell} \rangle = b_\ell$ for $\ell = 1, \ldots, m$, we consider the system of linear equations

$$\sum_{j=1}^{n} \langle \mathbf{v}_j, \mathbf{v}_{n_\ell} \rangle x_j = b_\ell = \langle \mathbf{v}, \mathbf{v}_{n_\ell} \rangle, \quad \ell = 1, \ldots, m, \qquad (3.3.5)$$

or in matrix formulation,

$$B\mathbf{x} = \mathbf{b}, \qquad (3.3.6)$$

where $\mathbf{x} = [x_1, \ldots, x_n]^T$ and

$$B = \left[\langle \mathbf{v}_j, \mathbf{v}_{n_\ell} \rangle \right], \qquad (3.3.7)$$

with $1 \leq \ell \leq m$ and $1 \leq j \leq n$, is the $m \times n$ coefficient matrix.

Therefore, by Theorem 1, the "solution" to (3.3.6) is given by

$$\mathbf{x}^\diamond = B^\dagger \mathbf{b}$$

(where $B^\dagger$ is the pseudo-inverse of B) in that

$$||B\mathbf{x}^\diamond - \mathbf{b}|| \leq ||B\mathbf{x} - \mathbf{b}||$$

for all $\mathbf{x} \in \mathbb{C}^n$ (or $\mathbf{x} \in \mathbb{R}^n$) and that $||\mathbf{x}^\diamond|| \leq ||\mathbf{y}||$ if

$$||B\mathbf{y} - \mathbf{b}|| = ||B\mathbf{x}^\diamond - \mathbf{b}||.$$

Of course, by setting $\mathbf{x}^\diamond = (x_1^\diamond, \ldots, x_n^\diamond)$, the (unique) optimal minimum-norm least-squares representation of $\mathbf{v} \in \mathbb{V}$ is given by

$$\sum_{j=1}^{n} x_j^\diamond \mathbf{v}_j. \tag{3.3.8}$$

Remark 1 Matching the inner product with the data vector $\mathbf{v}$ in (3.3.5) is a consequence of the variational method, when $\sum_j x_j \mathbf{v}_j$ is required to be the best approximation of $\mathbf{v}$ in the Euclidean (or ℓ_2) norm. Indeed, for the quantity $||\mathbf{v} - \sum_j x_j \mathbf{v}_j||^2$ to be the smallest for all choices of coefficients $x_1, \ldots, x_n$, the partial derivatives with respect to each $x_1, \ldots, x_n$ must be zero. For convenience, we only consider the real setting, so that for each $\ell = 1, \ldots, n$,

$$0 = \tfrac{\partial}{\partial x_\ell} ||\mathbf{v} - \sum_j x_j \mathbf{v}_j||^2$$

$$= \tfrac{\partial}{\partial x_\ell} \left(||\mathbf{v}||^2 - 2 \sum_{j=1}^{n} x_j \langle \mathbf{v}, \mathbf{v}_j \rangle + \sum_{j=1}^{n} \sum_{k=1}^{n} x_j x_k \langle \mathbf{v}_j, \mathbf{v}_k \rangle \right)$$

$$= -2 \langle \mathbf{v}, \mathbf{v}_\ell \rangle + \sum_{j=1}^{n} x_j \langle \mathbf{v}_j, \mathbf{v}_\ell \rangle + \sum_{k=1}^{n} x_k \langle \mathbf{v}_\ell, \mathbf{v}_k \rangle,$$

or

$$\sum_{j=1}^{n} x_j \langle \mathbf{v}_j, \mathbf{v}_\ell \rangle = \langle \mathbf{v}, \mathbf{v}_\ell \rangle, \tag{3.3.9}$$

which is (3.3.5), when n_ℓ is replaced by ℓ (see Exercise 9).

∎

Remark 2 For computational efficiency and stability, the coefficient matrix B in (3.3.6) should be "sparse" by choosing locally supported vectors (or functions) $\mathbf{v}_j$ in $\mathbb{V}$. For example, when piecewise polynomials (or splines) are used, it is best to use B-splines. In particular, when piecewise linear polynomials with equally spaced continuous "turning points" (called simple knots) are considered, then the linear B-splines are "hat" functions and the full matrix $B_h = [\langle v_j, v_k \rangle]$, for $1 \le j, \ k \le n$, is the banded square matrix

$$B_h = \frac{1}{6h} \begin{bmatrix} 2 & 1 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 1 & 4 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 1 & 4 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 1 & 4 & 1 \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 1 & 2 \end{bmatrix}, \tag{3.3.10}$$

where $h > 0$ is the distance between two adjacent knots (see Exercise 10).

∎

Exercises

Exercise 1 Show that the matrix $B^\dagger$ introduced in (3.4.1) satisfies (i)–(iv) in Theorem 2.

Exercise 2 In the proof of Theorem 2, show that the matrix sub-block $A_{12} = O$ by applying the assumption $(BA)^* = BA$ in (i).

Exercise 3 In the proof of Theorem 2, show that the matrix sub-block $A_{21} = O$ by applying the assumption $(AB)^* = AB$ in (ii).

Exercise 4 In the proof of Theorem 2, show that the matrix sub-block $A_{22} = O$ by applying the assumption $ABA = A$ in (iv).

Exercise 5 Compute the pseudo-inverse of each of the following matrices:

(a) $A_1 = \begin{bmatrix} 1 & 0 & 2 \\ 0 & 1 & 0 \end{bmatrix}$,

(b) $A_2 = \begin{bmatrix} -1 & 0 \\ 0 & 2 \\ 1 & 1 \end{bmatrix}$,

(c) $A_3 = \begin{bmatrix} -1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}$.

Exercise 6 Apply the results from Exercise 5 to find the general solutions of the following systems of linear equations. If there is more than one solution, then identify the solution with minimum ℓ_2-norm.

(a) $\begin{cases} x_1 + x_3 = 3, \\ x_2 = -1. \end{cases}$

(b) $\begin{cases} -x_1 = 2, \\ 2x_2 = 4, \\ x_1 + x_2 = 0. \end{cases}$

(c) $\begin{cases} -x_1 + x_2 = 0, \\ x_1 + x_3 = 5, \\ x_2 + x_3 = 5. \end{cases}$

Exercise 7 Apply the results from Exercise 5 to "solve" each of the following inconsistent systems of linear equations. Why is your "solution" reasonable?

(a) $\begin{cases} -x_1 = 2, \\ 2x_2 = 4, \\ x_1 + x_2 = 1. \end{cases}$

(b) $\begin{cases} -x_1 + x_2 = 0, \\ x_1 + x_3 = 5, \\ x_2 + x_3 = 1. \end{cases}$

Exercise 8 For any real number a, we introduce the notation $a_+ = \max(a, 0)$. Consider the function $B(x) = (1 - |x|)_+$ and its integer translates $B_j(x) = B(x - j)$. Show that $\{B_0(x), \ldots, B_n(x)\}$ constitutes a basis of the vector space $S_1[0, n]$ of continuous piecewise linear polynomials on $[0, n]$ with break-points (called knots) at the integers $1, \ldots, n-1$, by showing that

(a) every $f \in S_1[0, 1]$ has the formulation

$$f(x) = \sum_{k=0}^{n} f(k) B_j(x);$$

(b) $\{B_0(x), \ldots, B_n(x)\}$ is linearly independent.

Exercise 9 As a continuation of Exercise 8, show that for any function $g \in L_2[0, n]$,

$$\min_{x_1, \ldots, x_n \in \mathbb{R}} \left\| g(x) - \sum_{j=1}^{n} x_j B_j(x) \right\|_2$$

is achieved by

$$S_g(x) = \sum_{j=1}^{n} x_j B_j(x),$$

where $(x_0, \ldots, x_n)$ is the solution of the system of linear equations

$$\frac{\partial}{\partial x_\ell} \left\| g(x) - \sum_{j=1}^{n} x_j B_j(x) \right\|_2^2 = 0,$$

for $\ell = 0, \ldots, n$; or equivalently,

$$B \begin{bmatrix} x_0 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \langle g, B_0 \rangle \\ \vdots \\ \langle g, B_n \rangle \end{bmatrix},$$

where

$$B = \Big[\langle b_j, b_k \rangle \Big]_{0 \le j, \, k \le n},$$

and $\langle \cdot, \cdot \rangle$ denotes the inner product of $L_2[0, n]$.

Exercise 10 Compute the matrix B in the above Exercise 9.

Exercise 11 Derive the formulation of the coefficient matrix B_h for $v_j(x) = B(hx - j + 1)$ for $j = 1, \ldots, n + 1$ in (3.3.10) on p.154.

3.4 Data Dimensionality Reduction

The notion of principal components, introduced in Definition 2 on p.126 in Sect. 3.1, is instrumental to the study of data analysis. As mentioned in Remark 4 on p.128, the principal components provide a new coordinate system for the given data with the "hierarchal structure" in the decreasing order of the singular values for the data matrix. When applied to such problems as data management and data dimensionality reduction, current linear methods based on this data-dependent coordinate system are often collectively called "PCA", which stands for "principal component analysis".

On one hand, it is at least intuitively clear that effective data management facilitates the data reduction process, with the goal of eliminating the somewhat irrelevant data to reduce the data volume. In laymen's terms, perhaps "finding a needle in a haystack" pretty much describes the essence of "data mining". On the other hand, the notion of "data dimension" is certainly not a household word. For this reason, it is hoped that the following discussion will shed some light for understanding this important topic. When the dimension must be very high to represent the "data geometry", reduction of the data dimension is necessary (as long as the data geometry is not lost), for such applications as data feature extraction, data understanding, and data visualization.

So what is data dimension? Perhaps the most simple example to understand is color digital images, with data dimension equal to 3, for the primary color components R (red), G (green) and B (blue). Since human vision is "trichromatic", meaning that our retina contains three types of color cone cells with different absorption spectra, the three primary color components are combined, in various ratios, to yield a wide range of colors for human vision. However, with the rapid technological advancement in high-quality digital image and video display, more accurate color profiles for consistent imaging workflow require significantly more sophisticated color calibration, by using spectro-colorimeters that take narrow-band measurements, even below 10 nm (nano meter) increments. Hence, for the visible light (electromagnetic radiation, EMR) range of 400–700 nm, even a 10 nm increment requires 31 readings, a 10-fold increase over the 3 RGB readings. In other words, the "spectral curve" dimension for every single image pixel goes up from dimension 3 (for RGB) to dimension 31, and even higher if sub 10 nm increments are preferred. For many applications, including medical imaging, homeland security screening, satellite imaging for agriculture control and warfare, and so forth, the EMR range well beyond visible light is used. For example, a typical hyperspectral image (HSI), with 5 nm incremental reading on the EMR range between 320 nm (for ultraviolet) to 1600 nm (for infrared), already yields spectral curves in $\mathbb{R}^{257}$. Therefore, to facilitate computational efficiency, memory usage, data understanding and visualization, it is often necessary to reduce the data dimension while preserving data similarities (and dis-similarities), data geom-

etry, and data topology. In this elementary writing, we only discuss the linear PCA approach to dimensionality reduction.

Let B denote a dataset of m vectors $\mathbf{b}_1, \ldots, \mathbf{b}_m$ in $\mathbb{C}^n$. The objective for the problem of dimensionality reduction is to reduce n for the purpose of facilitating data understanding, analysis and visualization, without loss of the essential data information, such as data geometry and topology. For convenience, the notation for the set B is also used for the matrix $B \in \mathbb{C}^{m,n}$, called data-matrix, with row vectors:

$$\mathbf{b}_j^T = [b_{j,1}, \ldots, b_{j,n}],$$

or column vectors $\mathbf{b}_j$, where $j = 1, \ldots, m$; that is,

$$B = \begin{bmatrix} \mathbf{b}_1^T \\ \vdots \\ \mathbf{b}_m^T \end{bmatrix} = [\mathbf{b}_1 \ \cdots \ \mathbf{b}_m]^T.$$

Observe that the inner product of the jth and k^{th} rows of B, defined by

$$\langle \mathbf{b}_j, \mathbf{b}_k \rangle = \sum_{\ell=1}^n b_{j,\ell} \overline{b_{k,\ell}}, \tag{3.4.1}$$

reveals the "correlation" of the data $\mathbf{b}_j$ and $\mathbf{b}_k$ in terms of the ratio of its magnitude with respect to the product of their norms. Recall from the Cauchy-Schwarz inequality (1.4.4) on p.38 that, with $\| \cdot \| = \sqrt{\langle \cdot, \cdot \rangle}$,

$$|\langle \mathbf{b}_j, \mathbf{b}_k \rangle| \le \|\mathbf{b}_j\| \, \|\mathbf{b}_k\|,$$

and equality holds, if and only if $\mathbf{b}_k$ is a constant multiple of $\mathbf{b}_j$. Hence, $\mathbf{b}_j$ and $\mathbf{b}_k$ are a good "match" of each other if the ratio

$$\frac{|\langle \mathbf{b}_j, \mathbf{b}_k \rangle|}{\|\mathbf{b}_j\| \, \|\mathbf{b}_k\|} \tag{3.4.2}$$

(which is between 0 and 1) is close to 1, and a poor "match", if the ratio in (3.4.2) is close to 0. Since the inner product $\langle \mathbf{b}_j, \mathbf{b}_k \rangle$ is the $(j, k)^{\text{th}}$ entry of the $m \times m$ square matrix $A = BB^*$, called the Gram matrix of B, as defined by (3.1.1) on p.116, the Gram matrix of a dataset is often used to process the data.

On the other hand, even if $\mathbf{b}_j$ is only a minor "shift" of $\mathbf{b}_k$, the ratio in (3.4.2) could be much smaller than 1, which is achieved when $\mathbf{b}_j = \mathbf{b}_k$ (that is, without any shift). Hence, to provide a better measurement of data correlation, all data vectors are shifted by their average, as follows. For $\mathbf{b}_1, \ldots, \mathbf{b}_m \in \mathbb{C}^n$, their average is defined by

$$\mathbf{b}^{\text{av}} = \frac{1}{m} \sum_{j=1}^{m} \mathbf{b}_j, \tag{3.4.3}$$

and the **centered data-matrix** $\widetilde{B}$, associated with the given data-matrix B, is defined by

$$\widetilde{B} = \left[\widetilde{\mathbf{b}}_1, \ldots, \widetilde{\mathbf{b}}_m\right]^T = \left[\mathbf{b}_1 - \mathbf{b}^{\text{av}}, \cdots, \mathbf{b}_m - \mathbf{b}^{\text{av}}\right]^T. \tag{3.4.4}$$

We remark that the matrix

$$\frac{1}{m}\widetilde{B}(\widetilde{B})^* \tag{3.4.5}$$

is the **covariance matrix** of the dataset B, if $\mathbf{b}_1, \ldots, \mathbf{b}_m$ are observations of an n-dimensional random variable. Clearly, both Gram matrices $\widetilde{B}(\widetilde{B})^*$ and BB^* are spd matrices.

Again, when the same notation B of the dataset is used for the data-matrix $B = [\mathbf{b}_1 \ \cdots \ \mathbf{b}_m]^T$, the principal components of B, as introduced in Definition 2 on p.126, provide a new "coordinate system", with origin at $\mathbf{b}^{\text{av}}$ as defined in (3.4.3), which facilitates the analysis of the data B. All linear methods based on this data-dependent coordinate system are collectively called methods of **principal component analysis (PCA)**.

We remark that for PCA, because

$$\sum_{j=1}^{m} \widetilde{\mathbf{b}}_j = \sum_{j=1}^{m} (\mathbf{b}_j - \mathbf{b}^{\text{av}}) = \mathbf{0},$$

both the centered matrix $\widetilde{B}$ and its Gram $\widetilde{B}(\widetilde{B})^*$ are centered at $\mathbf{0}$, in the sense that the sum of all row vectors (of $\widetilde{B}$, and also of $\widetilde{B}\widetilde{B}^*$) is the zero vector $\mathbf{0}$. In addition, the geometry and topology of the data set B is unchanged, when B is replaced by $\widetilde{B}$, since for all $j, k = 1, \ldots, m$,

$$\|\widetilde{\mathbf{b}}_j - \widetilde{\mathbf{b}}_k\| = \|\mathbf{b}_j - \mathbf{b}_k\|. \tag{3.4.6}$$

In view of these nice properties of the centered matrix, when we say that PCA is applied to B, what we mean is that PCA is applied to the centered matrix $\widetilde{B}$ defined in (3.4.4).

To discuss PCA for data dimensionality reduction, let us first introduce the notion of "orthogonal rectangular matrices", as follows.

Definition 1 | **Orthogonal rectangular matrices** | *Let* $1 \leq d \leq q$ *be integers. The notation* $\mathcal{O}_{q,d}$ *is used for the collection of all* $q \times d$ *(complex) matrices* $W = [\mathbf{w}_1 \ \cdots \ \mathbf{w}_d]$ *with orthonormal column vectors; that is* $W^*W = I_d$ *or*

$$\langle \mathbf{w}_j, \mathbf{w}_k \rangle = \delta_{j-k}, \quad \text{all } 1 \leq j, k \leq d.$$

For any $W = [\mathbf{w}_1 \ \cdots \ \mathbf{w}_d] \in \mathcal{O}_{n,d}$, its column vectors constitute an orthonormal basis of its algebraic span,

$$\mathrm{span}\, W = \mathrm{span}\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_d\},$$

which is a d-dimensional subspace of $\mathbb{C}^n$, so that any $\mathbf{w} \in \mathrm{span}\, W$ has a unique representation:

$$\mathbf{w} = \sum_{j=1}^{d} c_j \mathbf{w}_j = W \begin{bmatrix} c_1 \\ \vdots \\ c_d \end{bmatrix},$$

for some $c_j \in \mathbb{C}$. The components of the vector

$$[c_1 \ \cdots \ c_d]^T$$

will be called the "coordinates" of vector $\mathbf{w}$, in the "coordinate system" with coordinate axes determined by the unit vectors: $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_d$. In this book, all coordinate systems are restricted to orthogonal systems, with orthonormal vectors as unit vectors for the coordinate axes.

We are now ready to study the topic of PCA-based dimensionality reduction. For the dataset $\{\mathbf{b}_1, \ldots, \mathbf{b}_m\} \subset \mathbb{C}^n$, consider the SVD of the data-matrix

$$B = [\mathbf{b}_1 \ \cdots \ \mathbf{b}_m]^T = USV^*, \tag{3.4.7}$$

for some unitary matrices $U \in \mathbb{C}^{m,m}$ and $V \in \mathbb{C}^{n,n}$ and an $m \times n$ matrix S, given by

$$S = \begin{bmatrix} \Sigma_r & O \\ O & O \end{bmatrix},$$

with $\Sigma_r = \mathrm{diag}\{\sigma_1, \ldots, \sigma_r\}$ and $\sigma_1 \geq \cdots \geq \sigma_r > 0$, where $r = \mathrm{rank}(B)$. Write U, V as

$$U = [\mathbf{u}_1 \ \cdots \ \mathbf{u}_m], \quad V = [\mathbf{v}_1 \ \cdots \ \mathbf{v}_n],$$

where $\mathbf{u}_j, \mathbf{v}_j$ are column vectors. Then the dimensionality reduction problem is to reduce $r = \mathrm{rank}(B)$ to an arbitrarily chosen integer d, with $1 \leq d < r$. The PCA approach is to consider the truncated matrices

$$U_d = [\mathbf{u}_1 \ \cdots \ \mathbf{u}_d], \quad V_d = [\mathbf{v}_1 \ \cdots \ \mathbf{v}_d], \quad \Sigma_d = \mathrm{diag}\{\sigma_1, \ldots, \sigma_d\}. \tag{3.4.8}$$

Then in view of the SVD $B = USV^*$ in (3.4.7), or equivalently $BV = US$, we consider the $m \times d$ matrix representation:

$$BV_d = U_d \Sigma_d,$$

and apply this $m \times d$ matrix $Y_d = BV_d = U_d\Sigma_d$ to introduce the notion of dimension-reduced data.

Definition 2 ⊡ **Dimension-reduced data** ⊡ *Let* $0 \le d < r$. *The column vectors* $\mathbf{y}_1, \ldots, \mathbf{y}_m$ *of the matrix* $Y_d^T = (BV_d)^T = (U_d\Sigma_d)^T$; *or equivalently,*

$$
\begin{bmatrix} \mathbf{y}_1^T \\ \vdots \\ \mathbf{y}_m^T \end{bmatrix} = Y_d = BV_d = U_d\Sigma_d, \tag{3.4.9}
$$

are said to constitute the dimension-reduced data of the given dataset $B = \{\mathbf{b}_1, \ldots, \mathbf{b}_m\} \subset \mathbb{C}^n$.

The main result on dimensionality reduction in this chapter is the following theorem which states that **the dimension-reduced dataset** $\mathbf{y}_1, \ldots, \mathbf{y}_m$ **of the given data** $B = \{\mathbf{b}_1, \ldots, \mathbf{b}_m\} \subset \mathbb{C}^n$ **is the optimal choice, in the sense that when measured in the "coordinate system"** $\overline{\mathbf{v}}_1, \ldots, \overline{\mathbf{v}}_d$ **in** $\mathbb{C}^n$, **this set provides the best** ℓ_2**-approximation of** B **among all possible choices** $\mathbf{q}_1, \ldots, \mathbf{q}_m \in \mathbb{C}^d$ **and all possible "coordinate systems"** W_d **of** $\mathbb{C}^d$.

Theorem 1 ⊡ **Dimension-reduced data are optimal** ⊡ *Let* $B = [\mathbf{b}_1 \ \cdots \ \mathbf{b}_m]^T$ *be an arbitrary* $m \times n$ *data-matrix with SVD representation given by* (3.4.7). *Then the dimension-reduced data* $\mathbf{y}_1, \ldots, \mathbf{y}_m$ *of* B, *defined by* (3.4.9), *lie in a d-dimensional subspace of* $\mathbb{C}^n$ *and satisfy the following best approximation property:*

$$
\sum_{j=1}^{m} \|\overline{V}_d\mathbf{y}_j - \mathbf{b}_j\|^2 \le \sum_{j=1}^{m} \|W_d\mathbf{q}_j - \mathbf{b}_j\|^2, \tag{3.4.10}
$$

for all $W_d = [\mathbf{w}_1 \ \cdots \ \mathbf{w}_d] \in \mathcal{O}_{n,d}$ *and all* $\{\mathbf{q}_1, \ldots, \mathbf{q}_m\} \subset \mathbb{C}^d$, *where* V_d, Σ_d *are defined by* (3.4.8) *with* $\overline{V}_d \in \mathcal{O}_{n,d}$.

In (3.4.10), for each $j = 1, \ldots, m$, the vector $\overline{V}_d\mathbf{y}_j$ lies in a d-dimensional subspace of $\mathbb{C}^n$ with basis $\overline{\mathbf{v}}_1, \ldots, \overline{\mathbf{v}}_d$, and the set $\overline{V}_d\mathbf{y}_1, \ldots, \overline{V}_d\mathbf{y}_m$ is an approximation of the given dataset $B = \{\mathbf{b}_1, \ldots, \mathbf{b}_m\}$. Observe that $\mathbf{y}_1, \ldots, \mathbf{y}_m$ are the coordinates of the approximants $\overline{V}_d\mathbf{y}_1, \ldots, \overline{V}_d\mathbf{y}_m$ in the coordinate system $\overline{\mathbf{v}}_1, \ldots, \overline{\mathbf{v}}_d$. Hence, the inequality (3.4.10) guarantees that the first d principal components $\mathbf{v}_1, \ldots, \mathbf{v}_d$ of B provide the best coordinate system (in hierarchal order), for optimal dimensionality reduction of the dataset $B \subset \mathbb{C}^n$ to a d-dimensional subspace, for any choice of dimension $d < n$. In particular, to reduce B to a 1-dimensional subspace, the first principal component $\mathbf{v}_1$ of B should be used to give the generator of this subspace for computing and representing the best 1-dimensional reduced data; to reduce B to a 2-dimensional subspace, the first and second principal components $\mathbf{v}_1, \mathbf{v}_2$ of B should be used for computing and representing the best 2-dimensional reduced data; and so forth. In general, to reduce the dimension of a given dataset $B = \{\mathbf{b}_1, \ldots, \mathbf{b}_m\} \subset \mathbb{C}^n$

to a d-dimensional subspace of $\mathbb{C}^n$, for any $d < n$, the best replacement of B is given by:

$$\begin{bmatrix} (\overline{V}_d \mathbf{y}_1)^T \\ \vdots \\ (\overline{V}_d \mathbf{y}_m)^T \end{bmatrix} = Y_d V_d^* = U_d \Sigma_d V_d^* = B V_d V_d^*. \tag{3.4.11}$$

Proof of Theorem 1   In view of (3.4.9) and the above dimension-reduced data representation formula (3.4.11), we may apply the formulation of $B(d)$ in (3.4.12) to write

$$\begin{bmatrix} (\overline{V}_d \mathbf{y}_1)^T \\ \vdots \\ (\overline{V}_d \mathbf{y}_m)^T \end{bmatrix} - B = Y_d V_d^* - B = [\mathbf{u}_1 \ \cdots \ \mathbf{u}_d] \Sigma_d [\mathbf{v}_1 \ \cdots \ \mathbf{v}_d]^* - B$$

$$= B(d) - B.$$

Hence, since the left-hand side of (3.4.10) can be written as $\sum_{j=1}^{m} \|\mathbf{y}_j^T V_d^* - \mathbf{b}_j^T\|^2$, which is precisely the square of the Frobenius norm of $Y_d V_d^* - B$ (see (3.2.8) on p.138), we have

$$\sum_{j=1}^{m} \|\overline{V}_d \mathbf{y}_j - \mathbf{b}_j\|^2 = \|B(d) - B\|_F^2.$$

Let Q be the $m \times d$ matrix with the jth row given by $\mathbf{q}_j^T$ for $1 \leq j \leq m$. Then the right-hand side of (3.4.10) can also be written as

$$\sum_{j=1}^{m} \|W_d \mathbf{q}_j - \mathbf{b}_j\|^2 = \sum_{j=1}^{m} \|\mathbf{q}_j^T W_d^T - \mathbf{b}_j^T\|^2 = \|R - B\|_F^2,$$

where

$$R = Q W_d^T. \tag{3.4.12}$$

Since $W_d \in \mathcal{O}_{n,d}$, the rank of the matrix R in (3.4.12) does not exceed d. Therefore, the desired inequality (3.4.10) follows from (3.4.14) in Theorem 7 on p.141. Furthermore, again by this theorem, (the square of) the error of the dimensionality reduction from $B \subset \mathbb{C}^n$ to $\mathbf{y}_1, \ldots, \mathbf{y}_m \subset \mathbb{C}^n$ is given by

$$\sum_{j=1}^{m} \|\overline{V}_d \mathbf{y}_j - \mathbf{b}_j\|^2 = \|B(d) - B\|_F^2 = \sum_{j=d+1}^{r} \sigma_j^2. \tag{3.4.13}$$

∎

Remark 1 The following comments should facilitate better understanding of the above PCA-based data reduction definition and result.

(i) Theorem 1 remains valid if the SVD of B in (3.4.7) is replaced by the reduced SVD

$$B = U_r \Sigma_r V_r^*, \qquad (3.4.14)$$

where U_r and V_r are $m \times r$ and $n \times r$ matrices, respectively, with orthonormal rows (that is, $U_r^* U_r = I_r$, $V_r^* V_r = I_r$), $r = $ rank (B), and Σ_r is the diagonal matrix of the (non-zero) singular values $\sigma_1 \geq \cdots \geq \sigma_r > 0$ of B.

(ii) If B has more rows than columns, it is more effective to compute the unitary matrix V for the SVD, $B = U S V^*$, from the spectral decomposition of $B^* B = V \Lambda V^*$ (instead of BB^*). In this undertaking, we may obtain the d-dimensional reduced data $\mathbf{y}_1, \ldots, \mathbf{y}_m$ from $Y_d = B V_d$ and the approximation of B by considering $B V_d V_d^*$ in (3.4.11).

(iii) Since the data set $B = \{\mathbf{b}_1 \quad \cdots \quad \mathbf{b}_m\}$ in $\mathbb{C}^n$ is not centered in general, the dimension-reduced data $\mathbf{y}_1, \ldots, \mathbf{y}_m$ of the centered dataset $\widetilde{B}$ (defined by (3.4.4), where $\mathbf{b}^{\mathrm{av}}$ is the average of B given by (3.4.3)), should be computed first. Then the required dimension-reduced dataset for the given dataset B is

$$\overline{V}_d \mathbf{y}_1 + \mathbf{b}^{\mathrm{av}}, \ldots, \overline{V}_d \, \mathbf{y}_m + \mathbf{b}^{\mathrm{av}},$$

and the error of approximation of $B = \{\mathbf{b}_1, \ldots, \mathbf{b}_m\}$ can be formulated as:

$$\left\| B V_d V_d^* + \begin{bmatrix} \mathbf{b}^{\mathrm{av}} \\ \vdots \\ \mathbf{b}^{\mathrm{av}} \end{bmatrix} - B \right\|_F \quad \text{or} \quad \left\| U_d \Sigma_d V_d^* + \begin{bmatrix} \mathbf{b}^{\mathrm{av}} \\ \vdots \\ \mathbf{b}^{\mathrm{av}} \end{bmatrix} - B \right\|_F.$$

Observe that the dimension-reduced data $\mathbf{y}_1, \ldots, \mathbf{y}_m$ lie in a d-dimensional subspace of $\mathbb{C}^n$, but of course, this subspace is usually not the same as $\mathbb{C}^d$.

(iv) In the above discussion, including (3.4.9) of the definition of dimension-reduced data and Theorem 1, we have only considered reduction of the data dimension n of the data set $B = \{\mathbf{b}_1, \ldots, \mathbf{b}_m\} \subset \mathbb{C}^n$. Analogous arguments (with appropriate new notation and definition) also apply to the problem of data reduction (in reducing the cardinality m of B), simply by applying the duality formulation of the SVD theorem:

$$B^* = V S^T U^*.$$

∎

Example 1 Let $B = [\mathbf{b}_1 \ \mathbf{b}_2 \ \mathbf{b}_3]^T \subset \mathbb{R}^2$, where

$$\mathbf{b}_1 = \begin{bmatrix} 2.5 \\ 4.5 \end{bmatrix}, \ \mathbf{b}_2 = \begin{bmatrix} 9 \\ 4 \end{bmatrix}, \ \mathbf{b}_3 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}.$$

Compute the dimension-reduced data $\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3\}$ of B, which lie in a 1-dimensional subspace of $\mathbb{R}^2$, by considering the dimension-reduced data $\{\tilde{\mathbf{y}}_1, \tilde{\mathbf{y}}_2, \tilde{\mathbf{y}}_3\}$ of the corresponding centered dataset $\tilde{B} = \left[\tilde{\mathbf{b}}_1; \tilde{\mathbf{b}}_2; \tilde{\mathbf{b}}_3\right]^T$ of B. In addition, compute the 2×1 (real) matrix transformation $V_1 = [\mathbf{v}_1]$ in (3.4.8), for which

$$\sum_{j=1}^{3} \|\tilde{\mathbf{y}}_j \mathbf{v}_1 - \tilde{\mathbf{b}}_j\|^2 \leq \sum_{j=1}^{3} \|q_j \mathbf{w}_1 - \tilde{\mathbf{b}}_j\|^2$$

for any unit vector $\mathbf{w}_1 \in \mathbb{R}^2$ and all $q_1, q_2, q_3 \in \mathbb{R}$.

Solution Since the average $\mathbf{b}^{\text{av}}$ of B is

$$\mathbf{b}^{\text{av}} = \begin{bmatrix} 3.5 \\ 2.5 \end{bmatrix},$$

we have

$$\tilde{\mathbf{b}}_1 = \mathbf{b}_1 - \mathbf{b}^{\text{av}} = \begin{bmatrix} -1 \\ 2 \end{bmatrix},$$

$$\tilde{\mathbf{b}}_2 = \mathbf{b}_2 - \mathbf{b}^{\text{av}} = \begin{bmatrix} 5.5 \\ 1.5 \end{bmatrix},$$

$$\tilde{\mathbf{b}}_3 = \mathbf{b}_3 - \mathbf{b}^{\text{av}} = \begin{bmatrix} -4.5 \\ -3.5 \end{bmatrix};$$

so that the centered dataset is given by

$$\tilde{B} = \begin{bmatrix} \tilde{\mathbf{b}}_1^T \\ \tilde{\mathbf{b}}_2^T \\ \tilde{\mathbf{b}}_3^T \end{bmatrix} = \begin{bmatrix} -1 & 2 \\ 5.5 & 1.5 \\ -4.5 & -3.5 \end{bmatrix}.$$

Since $\tilde{B}$ has more rows than columns, we may consider the spectral decomposition of $\tilde{B}^* \tilde{B}$ (instead of $\tilde{B} \tilde{B}^*$) to obtain V_1 in the reduced SVD of $\tilde{B}$ and apply $\tilde{Y}_1 = \tilde{B} V_1$ to obtain the reduced data for $\tilde{B}$, as follows. By direct calculation, we have

$$\tilde{B}^* \tilde{B} = \begin{bmatrix} 51.5 & 22 \\ 22 & 18.5 \end{bmatrix},$$

with eigenvalues $125/2$ and $15/2$ and corresponding eigenvectors given by

$$\mathbf{v}_1 = \begin{bmatrix} 2/\sqrt{5} \\ 1/\sqrt{5} \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} -1/\sqrt{5} \\ 2/\sqrt{5} \end{bmatrix}.$$

Hence, it follows from (3.4.9) that

$$Y_1 = \widetilde{B} V_1 = \widetilde{B} \mathbf{v}_1 = \begin{bmatrix} 0 \\ \frac{5\sqrt{5}}{2} \\ -\frac{5\sqrt{5}}{2} \end{bmatrix}.$$

That is, the dimension-reduced dataset for $\widetilde{B}$ is

$$\{\widetilde{y}_1, \widetilde{y}_2, \widetilde{y}_3\} = \left\{ 0, \frac{5\sqrt{5}}{2}, -\frac{5\sqrt{5}}{2} \right\}.$$

To verify that this solution is correct, we may compute (the square of) its error, as follows:

$$\sum_{j=1}^{3} \|\widetilde{y}_j \mathbf{v}_1 - \widetilde{\mathbf{b}}_j\|^2$$
$$= \|\widetilde{y}_1 \mathbf{v}_1 - \widetilde{\mathbf{b}}_1\|^2 + \|\widetilde{y}_2 \mathbf{v}_1 - \widetilde{\mathbf{b}}_2\|^2 + \|\widetilde{y}_3 \mathbf{v}_1 - \widetilde{\mathbf{b}}_3\|^2$$

$$= \|\mathbf{0} - \begin{bmatrix} -1 \\ 2 \end{bmatrix}\|^2 + \|\begin{bmatrix} 5 \\ 2.5 \end{bmatrix} - \begin{bmatrix} 5.5 \\ 1.5 \end{bmatrix}\|^2 + \| - \begin{bmatrix} 5 \\ 2.5 \end{bmatrix} + \begin{bmatrix} 4.5 \\ 3.5 \end{bmatrix}\|^2$$

$$= (1^2 + 2^2) + (0.5^2 + 1^2) + (0.5^2 + 1^2) = 7.5 = 15/2 = \sigma_2^2,$$

as expected from (3.4.13) in the proof of Theorem 1.

Hence, the dimension-reduced dataset $\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3$ of B can be obtained by adding $\mathbf{b}^{\mathrm{av}}$ to $\widetilde{y}_1 \mathbf{v}_1, \widetilde{y}_2 \mathbf{v}_1, \widetilde{y}_3 \mathbf{v}_1$, namely:

$$\mathbf{y}_1 = \begin{bmatrix} 3.5 \\ 2.5 \end{bmatrix}, \quad \mathbf{y}_2 = \begin{bmatrix} 8.5 \\ 5 \end{bmatrix}, \quad \mathbf{y}_3 = \begin{bmatrix} -1.5 \\ 0 \end{bmatrix}.$$

Observe that the three points $\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3$ lie on a straight line that passes through the point $\mathbf{b}^{\mathrm{av}} = (3.5, 2.5)$ with slope 0.5; and hence they are in a 1-dimensional space.

The centered dataset $\widetilde{B}$ is shown in the top box of Fig. 3.1; the dimension-reduced dataset $\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3$ of B (represented by 3 •), along with B, are shown in the bottom box of Fig. 3.1, while the first principal component $\mathbf{v}_1$ and second principal component $\mathbf{v}_2$ of B are also displayed in the bottom box of Fig. 3.1.

In the above discussion, the matrix $Y_d = \widetilde{B} V_d$ is applied to obtain the dimension-reduced data. If the reduced SVD of $\widetilde{B}$ is applied, we may use $Y_d = U_d \Sigma_d$ to obtain the reduced data instead. In the following, let us verify this fact. The singular values of $\widetilde{B}$ are computed by taking the square-roots of the eigenvalues of $\widetilde{B}^* \widetilde{B}$, namely: $\sigma_1 = \frac{5\sqrt{10}}{2}, \sigma_2 = \frac{\sqrt{30}}{2}$. The principal left eigenvectors $\mathbf{u}_1, \mathbf{u}_2$ can be obtained by applying (3.1.18) on p.120 with $r = 2$, as follows:

$$[\mathbf{u}_1 \ \mathbf{u}_2] = \widetilde{B} \ [\mathbf{v}_1 \ \mathbf{v}_2] \ \mathrm{diag}(\sigma_1^{-1}, \sigma_2^{-1}) = \begin{bmatrix} 0 & 2/\sqrt{6} \\ 1/\sqrt{2} & -1/\sqrt{6} \\ -1/\sqrt{2} & -1/\sqrt{6} \end{bmatrix},$$
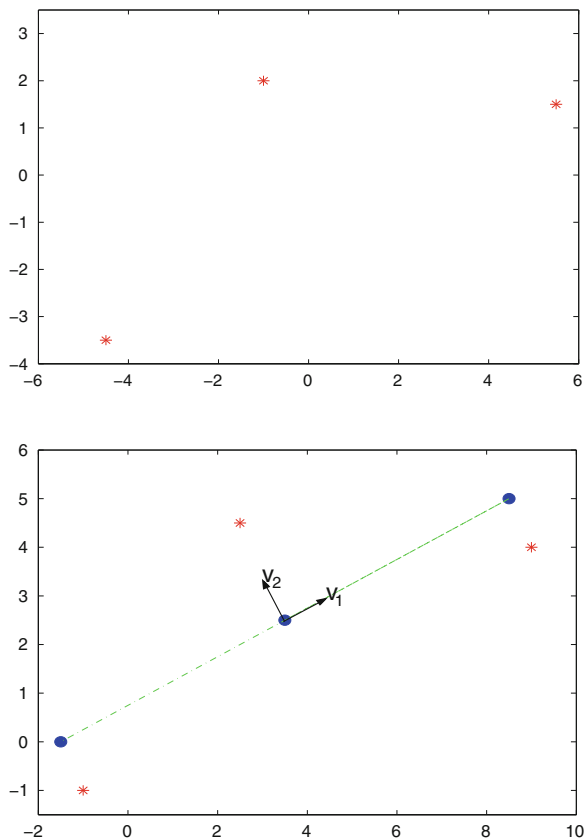
Fig. 3.1. *Top* Centered data (represented by $*$); *Bottom* original data (represented by $*$), dimension-reduced data (represented by $\bullet$), and principal components $\mathbf{v}_1$, $\mathbf{v}_2$

where the last equality is obtained by direct calculations. This enables us to compute

$$Y_1 = [\mathbf{u}_1][\sigma_1] = \frac{5\sqrt{10}}{2} \begin{bmatrix} 0 \\ \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} = \begin{bmatrix} 0 \\ \frac{5\sqrt{5}}{2} \\ -\frac{5\sqrt{5}}{2} \end{bmatrix},$$

which agrees with $Y_1 = \widetilde{B} V_1$ from the above computation. ∎

Example 2 Repeat Example 1 by considering the dataset matrix $B = [\mathbf{b}_1 \ \cdots \ \mathbf{b}_5]^T$ with 5 data points in $\mathbb{R}^3$ and reduce the dimension of the set to a 2-dimensional plane in $\mathbb{R}^3$, where
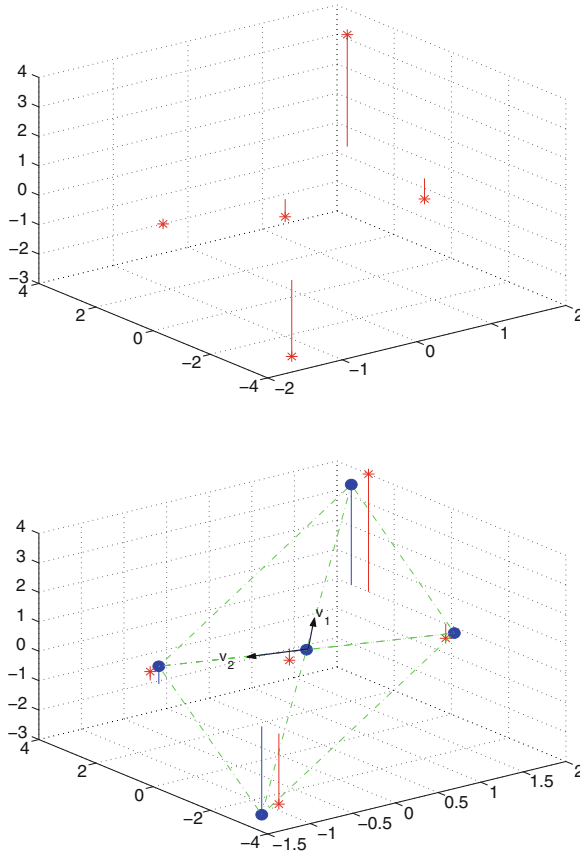
Fig. 3.2. *Top* centered data (represented by $*$); *bottom* dimension-reduced data (represented by $\bullet$) with the first and second principal components and original data (represented by $*$)

$$\mathbf{b}_1 = \begin{bmatrix} 2 \\ 2.9 \\ 4 \end{bmatrix}, \ \mathbf{b}_2 = \begin{bmatrix} -1.2 \\ 1 \\ 0.3 \end{bmatrix}, \ \mathbf{b}_3 = \begin{bmatrix} 0.4 \\ 0.9 \\ -0.4 \end{bmatrix}, \ \mathbf{b}_4 = \begin{bmatrix} 2 \\ 0.2 \\ -0.5 \end{bmatrix}, \ \mathbf{b}_5 = \begin{bmatrix} -1.2 \\ -3.5 \\ -2.4 \end{bmatrix}.$$

Solution Since the average $\mathbf{b}^{\mathrm{av}}$ of B is

$$\mathbf{b}^{\mathrm{av}} = [0.4, \ 0.3, \ 0.2]^T,$$

the matrix $\widetilde{B}$ of the centered dataset $\{\widetilde{\mathbf{b}}_1, \ \ldots, \ \widetilde{\mathbf{b}}_5\}$ is given by (see Fig. 3.2)

$$\widetilde{B} = \begin{bmatrix} 1.6 & 2.6 & 3.8 \\ -1.6 & 0.7 & 0.1 \\ 0 & 0.6 & -0.6 \\ 1.6 & -0.1 & -0.7 \\ -1.6 & -3.8 & -2.6 \end{bmatrix}.$$

To reduce computational cost, we compute the spectral decomposition of 3×3 matrix

$$\widetilde{B}^*\widetilde{B} = \frac{1}{50} \begin{bmatrix} 512 & 448 & 448 \\ 448 & 1103 & 977 \\ 448 & 977 & 1103 \end{bmatrix}$$

(instead of the 4×4 matrix $\widetilde{B}\widetilde{B}^*$). The eigenvalues of $\widetilde{B}^*\widetilde{B}$ are $1152/25$, $144/25$, and $63/25$, with the corresponding eigenvectors $\mathbf{v}_1$, $\mathbf{v}_2$ and $\mathbf{v}_3$, given by

$$\mathbf{v}_1 = \begin{bmatrix} 1/3 \\ 2/3 \\ 2/3 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} -4/(3\sqrt{2}) \\ 1/(3\sqrt{2}) \\ 1/(3\sqrt{2}) \end{bmatrix}, \quad \mathbf{v}_3 = \begin{bmatrix} 0 \\ 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix}.$$

Thus, it follows from (3.4.9) and Theorem 1 that

$$Y_2 = \widetilde{B}[\mathbf{v}_1 \ \mathbf{v}_2] = \begin{bmatrix} 24/5 & 0 \\ 0 & 6\sqrt{2}/5 \\ 0 & 0 \\ 0 & -6\sqrt{2}/5 \\ -24/5 & 0 \end{bmatrix}.$$

Therefore the dimension-reduced data for $\widetilde{\mathbf{b}}_1, \ldots, \widetilde{\mathbf{b}}_5$ are:

$$\begin{bmatrix} y_1 \\ z_1 \end{bmatrix} = \begin{bmatrix} \frac{24}{5} \\ 0 \end{bmatrix}, \quad \begin{bmatrix} y_2 \\ z_2 \end{bmatrix} = \begin{bmatrix} 0 \\ \frac{6\sqrt{2}}{5} \end{bmatrix}, \quad \begin{bmatrix} y_3 \\ z_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

$$\begin{bmatrix} y_4 \\ z_4 \end{bmatrix} = \begin{bmatrix} 0 \\ -\frac{6\sqrt{2}}{5} \end{bmatrix}, \quad \begin{bmatrix} y_5 \\ z_5 \end{bmatrix} = \begin{bmatrix} -\frac{24}{5} \\ 0 \end{bmatrix};$$

so that the dimension-reduced data for the given dataset B are $y_j\mathbf{v}_1 + z_j\mathbf{v}_2 + \mathbf{b}^{\mathrm{av}}$, for $j = 1, \ldots, 5$; namely,

$$\begin{bmatrix} 2 \\ 3.5 \\ 3.4 \end{bmatrix}, \quad \begin{bmatrix} -1.2 \\ 0.7 \\ 0.6 \end{bmatrix}, \quad \begin{bmatrix} 0.4 \\ 0.3 \\ 0.2 \end{bmatrix}, \quad \begin{bmatrix} 2 \\ -0.1 \\ -0.2 \end{bmatrix}, \quad \begin{bmatrix} -1.2 \\ -2.9 \\ -3 \end{bmatrix}.$$

These points lie on a 2-dimensional plane, generated by $\mathbf{v}_1$ and $\mathbf{v}_2$, that passes through the point $\mathbf{b}^{\mathrm{av}} = (0.4, 0.3, 0.2)$ as displayed in Fig. 3.2. ∎

<div align="center">

Exercises

</div>

Exercise 1 Consider the data points

$$\mathbf{b}_1 = \begin{bmatrix} 2 \\ -1 \end{bmatrix}, \quad \mathbf{b}_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \mathbf{b}_3 = \begin{bmatrix} 0 \\ -1 \end{bmatrix}, \quad \mathbf{b}_4 = \begin{bmatrix} 1 \\ 1 \end{bmatrix},$$

and let $\widetilde{B} = [\widetilde{\mathbf{b}}_1 \cdots \widetilde{\mathbf{b}}_4]^T$ be the matrix of the centered data $\widetilde{\mathbf{b}}_j = \mathbf{b}_j - \mathbf{b}^{\mathrm{av}}$, $1 \le j \le 4$, where $\mathbf{b}^{\mathrm{av}} = \frac{1}{4}(\mathbf{b}_1 + \cdots + \mathbf{b}_4)$. Compute the dimension-reduced data $\{y_1, y_2, y_3, y_4\} \subset \mathbb{R}$ of $\{\widetilde{\mathbf{b}}_1, \widetilde{\mathbf{b}}_2, \widetilde{\mathbf{b}}_3, \widetilde{\mathbf{b}}_4\}$, and determine the PCA-based dimension-reduced data $\mathbf{y}_1, \ldots, \mathbf{y}_4$ of $\mathbf{b}_1, \ldots, \mathbf{b}_4 \subset \mathbb{R}^2$ from $\mathbb{R}^2$ to a 1-dimensional subspace.

Hint: Follow the procedure in the solution of Example 1.

Exercise 2 Repeat Exercise 1 by considering

$$\mathbf{b}_1 = \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}, \quad \mathbf{b}_2 = \begin{bmatrix} -1 \\ 3 \\ 4 \end{bmatrix}, \quad \mathbf{b}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad \mathbf{b}_4 = \begin{bmatrix} 2 \\ 3 \\ -2 \end{bmatrix}.$$

Exercise 3 As in Exercise 2, let $\widetilde{B} = [\widetilde{\mathbf{b}}_1 \cdots \widetilde{\mathbf{b}}_4]^T$ be the centered data matrix of the data points

$$\mathbf{b}_1 = \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix}, \quad \mathbf{b}_2 = \begin{bmatrix} -1 \\ 2 \\ 3 \end{bmatrix}, \quad \mathbf{b}_3 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad \mathbf{b}_4 = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}.$$

Compute the dimension-reduced set $\{\widetilde{\mathbf{y}}_1, \widetilde{\mathbf{y}}_2, \widetilde{\mathbf{y}}_3, \widetilde{\mathbf{y}}_4\}$ with $\widetilde{\mathbf{y}}_j = [\widetilde{y}_j, \widetilde{z}_j]^T \in \mathbb{R}^2$ of $\{\widetilde{\mathbf{b}}_1, \widetilde{\mathbf{b}}_2, \widetilde{\mathbf{b}}_3, \widetilde{\mathbf{b}}_4\}$, and determine the PCA-based dimension-reduced data $\{\mathbf{y}_1, \ldots, \mathbf{y}_4\}$ of $\{\mathbf{b}_1, \ldots, \mathbf{b}_4\} \subset \mathbb{R}^3$ from $\mathbb{R}^3$ to a 2-dimensional subspace.

Hint: Follow the procedure in the solution of Example 2.

Chapter 4
Frequency-Domain Methods
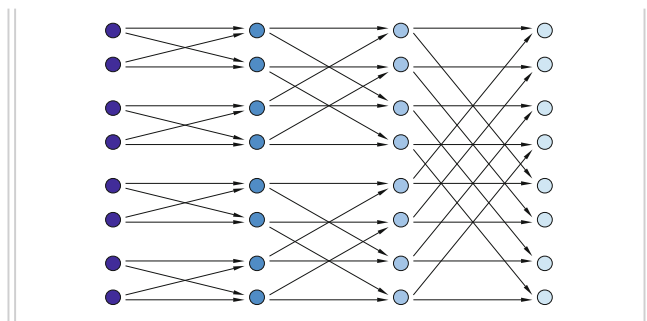
Spectral methods based on the singular value decomposition (SVD) of the data matrix, as studied in the previous chapter, Chap. 3, apply to the physical (or spatial) domain of the data set. In this chapter, the concepts of frequency and frequency representation are introduced, and various frequency-domain methods along with efficient computational algorithms are derived. The root of these methods is the Fourier series representation of functions on a bounded interval, which is one of the most important topics in Applied Mathematics and will be investigated in some depth in Chap. 6. While the theory and methods of Fourier series require knowledge of mathematical analysis, the discrete version of the Fourier coefficients (of the Fourier series) is simply matrix multiplication of the data vector by some $n \times n$ square matrix F_n. This matrix is called the discrete Fourier transformation (DFT), which has the important property that with the multiplicative factor of $\frac{1}{\sqrt{n}}$, it becomes a unitary matrix, so that the inverse discrete Fourier transformation matrix (IDFT) is given by the $\frac{1}{n}$- multiple of the adjoint (that is, transpose of the complex conjugate) of F_n. This topic will be studied in Sect. 4.1.

For the investigation of the frequency contents of n-dimensional real-valued data vectors, it is certainly more effective to apply some real-valued version of the DFT. In Sect. 4.2, the notion of discrete cosine transform (DCT), to be denoted by C_n, is introduced; and the most popular version, DCT-II, which is one of the four types of DCT's, is derived from the DFT matrix, F_{2n}, introduced in Sect. 4.1. The formulation of this DCT includes the $\frac{1}{\sqrt{n}}$ multiplicative factor, and is therefore an orthogonal (that is, a real-valued unitary) matrix, with the corresponding inverse DCT (IDCT) given by the transpose of C_n. The reason for the choice of DCT, as opposed to an analogous discrete sine transform (DST), for the transformation of data vectors from the spatial domain to the frequency domain, is that DCT trivially "encodes" the zero-frequency contents of the data vectors, while the computational cost for encoding the zero-frequency contents by DST is very high.

Since the computational complexity of the n-point DFT is of order $O(n^2)$, which is unacceptably high for large values of n, the notion of fast Fourier transform (FFT) is introduced in Sect. 4.3 via Lanczos' matrix factorization formula for F_n provided that n is an even integer, in order to decrease the complexity count. This will be Theorem 1 of the section. Consequently, a full factorization formula of the DFT, F_n for $n = 2^m$, can be established and will be stated in Theorem 2, to reduce the computation count of the DFT, F_{2^m}, to $O(nm) = O(n \log n)$ arithmetic operations. Of course this count is meaningless for small values of $n > 0$. For this reason, precise counts are derived for $n = 4, 8, 16$ in this section also. The reason for inclusion of F_{16} is that it is used to formulate the DCT-II matrix, C_8, as discussed in Sect. 4.2. In addition, signal flow charts are also included at the end of this section. In Sect. 4.4, the other three types of DCT's, namely: DCT-I, DCT-III, and DCT-IV, are formulated. In addition, Theorem 2 of Sect. 4.3 is applied to formulate the fast DCT (FDCT) for all four types of DCT's. This final section of Chap. 4 ends with the display of various data flow charts for FDCT implementation of the 8-point DCT-II, which is used for both JPEG digital image, as well as MPEG video, compressions, to be discussed in Sect. 5.4 of the next chapter.

4.1 Discrete Fourier Transform

In view of the invariance properties of unitary matrices, as described in Sect. 2.3, unitary matrices are often used in applications, such as transformation of a digital signal (from the physical domain) to the "frequency domain".

The notion of "frequency" is associated with that of "wavelength", assuming that the digital signal is obtained by discretization (that is, sampling followed by quantization) of an analog signal in the form of a traveling wave (since the analogous concept of standing waves is not discussed in this context). There are various definitions of wavelengths that are somewhat equivalent. In this book, the distance (on the time axis) between two consecutive local maxima of the waveform is used as the definition of wavelength, say, η m (where m stands for "meters"). Let us assume that the wave under consideration travels at the speed of v_0 m/s (which stands for "meters per second"), then the frequency is defined by

$$\omega_0 = \frac{v_0}{\eta} \text{ Hz,} \tag{4.1.1}$$

where 1 Hz (which stands for "Hertz") is one cycle per second. Typical examples include electro-magnetic waves that travel at the speed of light, namely, $c = 299,792,458$ m/s (in vacuum) and sound waves, whose speed depends not only on the medium of wave propagation, but also on the temperature. For a repeating event, the frequency is the reciprocal of the period.

Mathematical formulations of traveling waves depend on sinusoidal functions, in terms of cosines and/or sines. For example, a single-frequency signal represented as

$$w(t) \;=\; a \cos 2\pi\omega t + b \sin 2\pi\omega t$$

for some $a, b \in \mathbb{R}$, has the frequency $= \omega$ Hz, if the unit of measurement in the time-domain t is in seconds. Since both $\cos 2\pi\omega t$ and $\sin 2\pi\omega t$ can be formulated in terms of $e^{i2\pi\omega t}$ and $e^{-i2\pi\omega t}$, where $i = \sqrt{-1}$ is the imaginary unit, via "Euler's identity"

$$e^{ix} \;=\; \cos x + i \sin x,$$

namely:

$$\cos x = \frac{e^{ix} + e^{-ix}}{2};$$

$$\sin x = \frac{e^{ix} - e^{-ix}}{2i},$$

it is important to acquire the basic skill for manipulating the function e^{ix}. The following three formulas are particularly useful.

(i) Geometric sum:

$$1 + e^{ix} + \cdots + e^{i(n-1)x} = \frac{1 - e^{inx}}{1 - e^{ix}}. \tag{4.1.2}$$

(ii) Orthonormal sequences:

$$\frac{1}{n} \sum_{k=0}^{n-1} e^{i2\pi k(\ell - j)/n} = \delta_{\ell - j}, \quad 0 \le \ell, \ j \le n - 1. \tag{4.1.3}$$

In other words, by introducing the set $\{\mathbf{f}_\ell\}$, $\ell = 0, \ldots, n-1$, of sequences $\mathbf{f}_\ell$, defined by

$$\mathbf{f}_\ell = (a_{\ell,0}, \ldots, a_{\ell,n-1}), \ \ a_{\ell,k} = e^{i(2\pi\ell/n)k}, \tag{4.1.4}$$

we have

$$\ll \mathbf{f}_\ell, \mathbf{f}_j \gg = \delta_{\ell - j}, \tag{4.1.5}$$

where $\ll, \gg$ is the normalized inner product defined by

$$\ll \mathbf{f}_\ell, \mathbf{f}_j \gg = \frac{1}{n} \sum_{k=0}^{n-1} a_{\ell,k} \bar{a}_{k,j}. \tag{4.1.6}$$

(iii) Orthonormal basis functions:

$$\frac{1}{2\pi} \int_0^{2\pi} e^{i(\ell - j)x} dx = \delta_{\ell - j}, \tag{4.1.7}$$

for all integers ℓ and j. In other words, the family $\{g_\ell\}$ of functions g_ℓ, where ℓ runs over all integers, defined by

$$g_\ell(x) = e^{i\ell x}, \quad 0 \le x \le 2\pi,$$

is an orthonormal family with respect to the normalized inner product $\langle\!\langle \ , \ \rangle\!\rangle$

$$\langle\!\langle g_\ell, g_j \rangle\!\rangle = \frac{1}{2\pi} \int_0^{2\pi} g_\ell(x)\overline{g_j(x)}dx = \delta_{\ell-j}.$$

Definition 1 | **Discrete Fourier transformation (DFT)** | *Let $n \ge 2$ be any given integer. Then the n-point* **discrete Fourier transformation** *(DFT) is the $n \times n$ matrix F_n of complex numbers, defined by*

$$F_n = \left[(z_n)^{jk}\right]_{0 \le j,k \le n-1} = \begin{bmatrix} 1 & z_n^0 & \cdots & z_n^0 \\ 1 & z_n^1 & \cdots & z_n^{n-1} \\ & & \cdots & \\ 1 & z_n^{n-1} & \cdots & z_n^{(n-1)(n-1)} \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & z_n & \cdots & z_n^{n-1} \\ & & \cdots & \\ 1 & z_n^{n-1} & \cdots & z_n^{(n-1)(n-1)} \end{bmatrix}, \tag{4.1.8}$$

where

$$z_n = e^{-i2\pi/n}. \tag{4.1.9}$$

Furthermore, the n-point **inverse discrete Fourier transformation** *(IDFT) is defined by $\widetilde{F}_n = \frac{1}{n}F_n^*$, where F_n^* is the adjoint (that is, transpose and complex conjugate) of F_n.*

For a vector $\mathbf{x} \in \mathbb{C}^n$, its **discrete Fourier transform (DFT)**, denoted by $\widehat{\mathbf{x}}$, is defined as

$$\widehat{\mathbf{x}} = F_n\mathbf{x}.$$

Theorem 1 *For each $n \ge 2$, let F_n be defined as in* (4.1.8)–(4.1.9). *Then the matrix U_n, defined by*

$$U_n = \frac{1}{\sqrt{n}}F_n, \tag{4.1.10}$$

is a unitary matrix. Hence,

$$F_n\widetilde{F}_n = \widetilde{F}_n F_n = I_n, \tag{4.1.11}$$

where $I_n = \mathrm{diag}\{1, \ldots, 1\}$ is the identity matrix.

Since (4.1.11) immediately follows from (4.1.10) and the definition of $\widetilde{F}_n$, it is sufficient to verify (4.1.10).

To verify the validity of (4.1.10), observe, according to (4.1.8)−(4.1.9) and (4.1.4), that

$$F_n = \begin{bmatrix} a_{0,0} & a_{1,0} & \cdots & a_{n-1,0} \\ \vdots & \vdots & \vdots & \vdots \\ a_{0,n-1} & a_{1,n-1} & \cdots & a_{n-1,n-1} \end{bmatrix}.$$

Hence, for $0 \leq \ell, j \leq n - 1$, the (ℓ, j)th entry of the matrix UU^* is precisely

$$\sum_{k=0}^{n-1} \left(\frac{1}{\sqrt{n}} a_{\ell,k}\right) \overline{\left(\frac{1}{\sqrt{n}} a_{k,j}\right)} = \frac{1}{n} \sum_{k=0}^{n-1} a_{\ell,k} \bar{a}_{k,j} = \delta_{\ell-j},$$

where the last equality follows from (4.1.6) and (4.1.5) consecutively. That is, we have shown that $UU^* = I_n$. ∎

Example 1 Formulate and simplify the 4-point DFT and IDFT (matrix) operators, F_4 and $\widetilde{F}_4$, respectively. Verify (4.1.11) for $n = 4$; that is,

$$F_4 \widetilde{F}_4 = I_4.$$

Solution For $n = 4$, we have $z_4 = e^{-i2\pi/4} = e^{-i\pi/2} = \cos(-\frac{\pi}{2}) + i \sin(-\frac{\pi}{2}) = -i$. Hence, it follows from (4.1.8) that

$$F_4 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & (-i) & (-i)^2 & (-i)^3 \\ 1 & (-i)^2 & (-i)^4 & (-i)^6 \\ 1 & (-i)^3 & (-i)^6 & (-i)^9 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -i & -1 & i \\ 1 & -1 & 1 & -1 \\ 1 & i & -1 & -i \end{bmatrix}.$$

In addition, by the definition of IDFT, we have

$$\widetilde{F}_4 = \frac{1}{4} F^* = \frac{1}{4} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & i & -1 & -i \\ 1 & -1 & 1 & -1 \\ 1 & -i & -1 & i \end{bmatrix}^T$$

$$= \frac{1}{4} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & i & -1 & -i \\ 1 & -1 & 1 & -1 \\ 1 & -i & -1 & i \end{bmatrix}.$$

Finally, to verify (4.1.11) for $n = 4$, we simply compute

$$
F_4 \widetilde{F}_4 = \frac{1}{4}
\begin{bmatrix}
1 & 1 & 1 & 1 \\
1 & -i & -1 & i \\
1 & -1 & 1 & -1 \\
1 & i & -1 & -i
\end{bmatrix}
\begin{bmatrix}
1 & 1 & 1 & 1 \\
1 & i & -1 & -i \\
1 & -1 & 1 & -1 \\
1 & -i & -1 & i
\end{bmatrix}
$$

$$
= \frac{1}{4}
\begin{bmatrix}
4 & (1+i-1-i) & (1-1+1-1) & (1-i-1+i) \\
(1-i-1+i) & (1+1+1+1) & (1+i-1-i) & (1-1+1-1) \\
(1-1+1-1) & (1-i-1+i) & (1+1+1+1) & (1+i-1-i) \\
(1+i-1-i) & (1-1+1-1) & (1-i-1+i) & (1+1+1+1)
\end{bmatrix}
$$

$$
= \frac{1}{4}
\begin{bmatrix}
4 & 0 & 0 & 0 \\
0 & 4 & 0 & 0 \\
0 & 0 & 4 & 0 \\
0 & 0 & 0 & 4
\end{bmatrix}
=
\begin{bmatrix}
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1
\end{bmatrix}.
$$

∎

Example 2 Apply F_4 from Example 1 to compute the DFT of the following digital signals (formulated as column vectors for convenience).

(i) $\mathbf{x} = [1, 1, 1, 1]^T$.
(ii) $\mathbf{y} = [0, 1, 2, 3]^T$.
(iii) $\mathbf{z} = [0, 1, 0, 1]^T$.

Interpret the frequency content of $\widehat{\mathbf{x}} = F_4\mathbf{x}$, $\widehat{\mathbf{y}} = F_4\mathbf{y}$, and $\widehat{\mathbf{z}} = F_4\mathbf{z}$.

Solution For (i),

$$
\widehat{\mathbf{x}} = F_4\mathbf{x} =
\begin{bmatrix}
1 & 1 & 1 & 1 \\
1 & -i & -1 & i \\
1 & -1 & 1 & -1 \\
1 & i & -1 & -i
\end{bmatrix}
\begin{bmatrix}
1 \\
1 \\
1 \\
1
\end{bmatrix}
=
\begin{bmatrix}
4 \\
0 \\
0 \\
0
\end{bmatrix}.
$$

Since the signal $\mathbf{x}$ is a constant, there is no high-frequency content. Hence, the DC (direct current) term is 4 and the AC (alternate current) contents are $0, 0, 0$.

For (ii),

$$
\widehat{\mathbf{y}} = F_4\mathbf{y} =
\begin{bmatrix}
1 & 1 & 1 & 1 \\
1 & -i & -1 & i \\
1 & -1 & 1 & -1 \\
1 & i & -1 & -i
\end{bmatrix}
\begin{bmatrix}
0 \\
1 \\
2 \\
3
\end{bmatrix}
$$

$$
=
\begin{bmatrix}
6 \\
2(-1+i) \\
-2 \\
2(-1-i)
\end{bmatrix}
=
\begin{bmatrix}
6 \\
-2 \\
-2 \\
-2
\end{bmatrix}
+ i
\begin{bmatrix}
0 \\
2 \\
0 \\
-2
\end{bmatrix}.
$$

Since the signal $\mathbf{y}$ does not oscillate, the DC term ($= 6$) is the largest. Observe that both the real and imaginary parts of the AC terms, being $[-2, -2, -2]$ and $[2, 0, -2]$, have constant slopes (0 and -2, respectively). Hence, the "phase" of $\widehat{\mathbf{y}} = F_4 \mathbf{y}$ is linear.

For (iii),

$$\widehat{\mathbf{z}} = F_4 \mathbf{z} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -i & -1 & i \\ 1 & -1 & 1 & -1 \\ 1 & i & -1 & -i \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \\ -2 \\ 0 \end{bmatrix}.$$

The DC term is small (only $= 2$), while only one high-frequency term is non-zero. Indeed, the signal $(0, 1, 0, 1)$ has only one single period. ∎

Example 3 Let S_1, $S_2 \in \mathbb{R}^{64}$ be signals given by

$$S_1(j) = \begin{cases} 1, & 0 \le j \le 15, \\ 2, & 16 \le j \le 31, \\ -1, & 32 \le j \le 47, \\ 0.5, & 48 \le j \le 63; \end{cases}$$

and

$$S_2(j) = j/8, \quad 0 \le j \le 63.$$

The real part and imaginary part of the DFT of S_1 are shown in Fig. 4.1, those of the DFT of S_2 are shown in Fig. 4.2. ∎

By using the column vector notation for the finite sequences $\mathbf{f}_0, \ldots, \mathbf{f}_{n-1}$ introduced in (4.1.4), we may formulate the n-point DFT of any $\mathbf{x} = [x_0, \ldots, x_{n-1}]^T \in \mathbb{C}$ as follows:

$$\widehat{\mathbf{x}} = \begin{bmatrix} \widehat{x}_0 \\ \vdots \\ \widehat{x}_{n-1} \end{bmatrix} = F_n \mathbf{x} = \begin{bmatrix} \langle \mathbf{x}, \mathbf{f}_0 \rangle \\ \vdots \\ \langle \mathbf{x}, \mathbf{f}_{n-1} \rangle \end{bmatrix}.$$

Hence, to compute the DFT $\widehat{\mathbf{x}}$ of $\mathbf{x} \in \mathbb{C}^n$, it is necessary to compute each of the n inner products:

$$\langle \mathbf{x}, \mathbf{f}_\ell \rangle = \sum_{k=0}^{n-1} x_k e^{-i2\pi k\ell/n}, \quad \ell = 0, \ldots, n-1. \tag{4.1.12}$$

Let us assume that $\cos(2\pi k\ell/n)$ and $\sin(2\pi k\ell/n)$ have been pre-computed and stored in a look-up table. Then the number of multiplications in (4.1.12) is $2(n-1)$. Taking into account of n additions in (4.1.12), the number of operations required to compute $\langle \mathbf{x}, \mathbf{f}_\ell \rangle$ is $3n - 2$; so that the total number of operations to compute the DFT $\widehat{\mathbf{x}}$ of $\mathbf{x}$ is $(3n^2 - 2n)$. Hence, the amount of computations is quite intensive for large n.
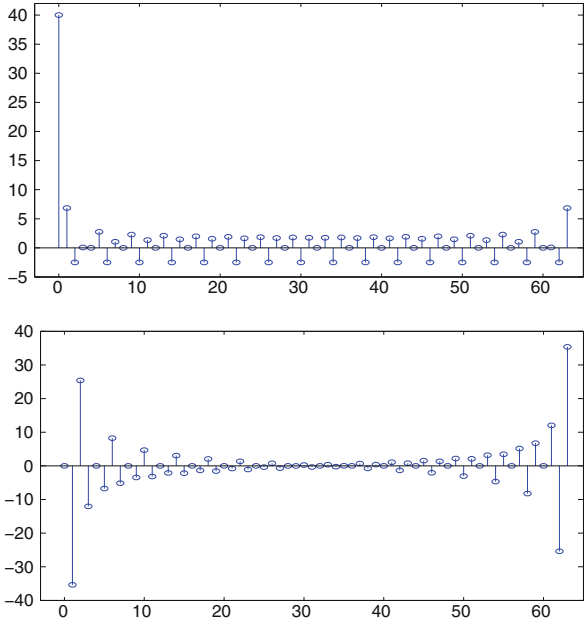
Fig. 4.1 *Top* real part of the DFT of S_1; *Bottom* imaginary part of the DFT of S_1
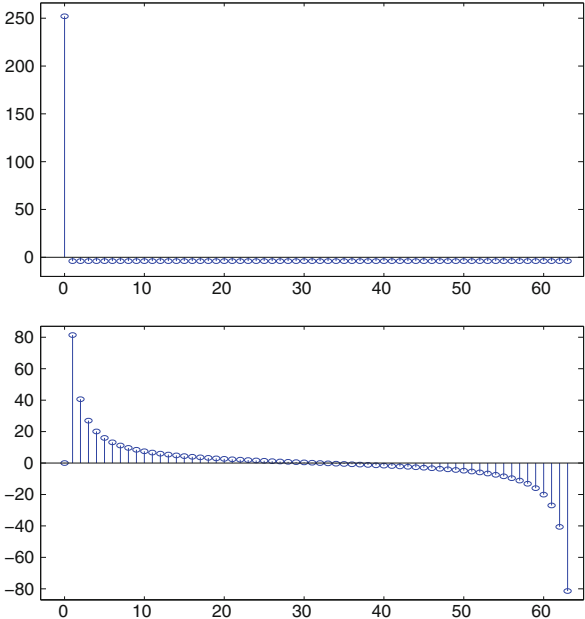
Fig. 4.2 *Top* real part of the DFT of S_2; *Bottom* imaginary part of the DFT of S_2

The fast Fourier transform (FFT) to be discussed in a later section reduces the order of operations from $O(n^2)$ to $O(n \log n)$, for $n = 2^m$, $m = 1, 2, \ldots$.

<div align="center">

| **Exercises** |

</div>

Exercise 1 Compute the frequency for each of the following electromagnetic (EM) waves $w_j(t)$, $j = 1, 2, 3, 4$, with given wavelengths η_j. (*Hint*: EM waves travel at the speed of light. Also 1 nm (nanometer) $= 10^{-9}$ m (meter)).

(a) $w_1(t)$ with $\eta_1 = 600$ nm (visible green light).
(b) $w_2(t)$ with $\eta_2 = 350$ nm (ultraviolet).
(c) $w_3(t)$ with $\eta_3 = 1,999$ nm (infrared).
(d) $w_4(t)$ with $\eta_4 = 0.1$ nm (X-ray).

Exercise 2 Write down the following complex numbers in the form of $a + ib$, where a and b are real numbers. Simplify your answers.

(a) $e^{i2\pi k/3}$, $k = 0, 1, 2$.
(b) $e^{i2\pi k/4}$, $k = 0, 1, 2, 3$.
(c) $e^{i2\pi k/6}$, $k = 0, 1, 2, 3, 4, 5$.
(d) $e^{-i\pi k/2}$, $k = 0, 1, 2, 3$.

Exercise 3 Fill in the details in the derivation of (4.1.2), (4.1.3), (4.1.6), and (4.1.7).

Exercise 4 Apply the results from Exercise 2 to write down the DFT and IDFT matrices F_3, $\widetilde{F}_3$, F_4, $\widetilde{F}_4$, F_6, $\widetilde{F}_6$, with precise (not approximate) values of all entries in the form of $a + ib$, where $a, b \in \mathbb{R}$. (Hint: See Example 1 for F_4 and $\widetilde{F}_4$.)

Exercise 5 Compute the DFT of the given vectors $\mathbf{x}$ and $\mathbf{y}$. Verify your answers by computing the corresponding IDFT.

(a) $F_3\mathbf{x}$ and $F_3\mathbf{y}$, for $\mathbf{x} = [1, -1, 1]^T$ and $\mathbf{y} = [i, 0, -i]^T$.
(b) $F_4\mathbf{x}$ and $F_4\mathbf{y}$, for $\mathbf{x} = [0, i, 0, -i]^T$ and $\mathbf{y} = [1, 0, -1, 1]^T$.

Exercise 6 Give explicit formulation of the 8-point DFT matrix F_8 in terms of F_4 by applying Theorem 1 and the matrix F_4 in Example 1.

Exercise 7 Repeat Exercise 6 for the 6-point DFT matrix F_6 in terms of F_3 from Exercise 4.

4.2 Discrete Cosine Transform

To introduce the discrete cosine transform (DCT), let us recall the DFT in Sect. 4.1, namely the n-point DFT in (4.1.8)–(4.1.9) of Definition 1 on p.174. In applications

to audio, image, and video compression, since complex numbers in DFT are more costly to code (or more precisely to "encode"), the real DCT is preferred.

Let $\mathbf{x} = [x_0, \ldots, x_{n-1}]^T \in \mathbb{R}^n$, and define

$$\widetilde{x}_\ell = \begin{cases} x_\ell, & \text{for } \ell = 0, \ldots, n-1, \\ x_{2n-\ell-1}, & \text{for } \ell = n, \ldots, 2n-1, \end{cases} \tag{4.2.1}$$

to extend the vector $\mathbf{x}$ from $\mathbb{R}^n$ to the vector $\widetilde{\mathbf{x}} = [\widetilde{x}_0, \ldots, \widetilde{x}_{2n-1}]$ in $\mathbb{R}^{2n}$, such that the extended vector $\widetilde{\mathbf{x}}$ is symmetric with respect to the center "$n - \frac{1}{2}$", as follows:

$$\widetilde{x}_n = \widetilde{x}_{n-1} = x_{n-1}$$
$$\widetilde{x}_{n+1} = \widetilde{x}_{n-2} = x_{n-2}$$
$$\widetilde{x}_{n+2} = \widetilde{x}_{n-3} = x_{n-3}$$
$$\ldots\ldots\ldots\ldots\ldots$$
$$\widetilde{x}_{2n-1} = \widetilde{x}_0 = x_0.$$

Now, apply (4.1.8) with n replaced by $2n$ to formulate the $2n$-point DFT $\widehat{\widetilde{\mathbf{x}}}$ of $\widetilde{\mathbf{x}}$, with jth component (of $\widehat{\widetilde{\mathbf{x}}}$) given by

$$\left(\widehat{\widetilde{\mathbf{x}}}\right)_j = \sum_{\ell=0}^{2n-1} \widetilde{x}_\ell e^{-i2\ell\pi \frac{j}{2n}}$$

$$= \sum_{\ell=0}^{n-1} x_\ell e^{-i\frac{\ell j\pi}{n}} + \sum_{\ell=n}^{2n-1} x_{2n-\ell-1} e^{-i\frac{\ell j\pi}{n}} \tag{4.2.2}$$

for $j = 0, \ldots, 2n-1$, where the definition of $\widetilde{x}_\ell$ in (4.2.1) is used. But the last summation in (4.2.2) can be simplified to be

$$\sum_{\ell=n}^{2n-1} x_{2n-\ell-1} e^{\frac{-i\ell j\pi}{n}} = \sum_{\ell=0}^{n-1} x_\ell e^{\frac{i(\ell+1)j\pi}{n}},$$

since $e^{-i2n\pi/n} = e^{-i2\pi} = 1$. Hence, putting this back into (4.2.2) yields

$$\left(\widehat{\widetilde{\mathbf{x}}}\right)_j = \sum_{\ell=0}^{n-1} x_\ell \left(e^{-i\frac{\ell j\pi}{n}} + e^{i\frac{\ell j\pi}{n}} e^{i\frac{j\pi}{n}}\right)$$

$$= e^{\frac{ij\pi}{2n}} \sum_{\ell=0}^{n-1} x_\ell \left(e^{-ij(\frac{2\ell+1}{2n})\pi} + e^{ij(\frac{2\ell+1}{2n})\pi}\right)$$

$$= 2e^{\frac{ij\pi}{2n}} \sum_{\ell=0}^{n-1} x_\ell \cos\frac{j(2\ell+1)\pi}{2n} = d_j e^{ij\frac{\pi}{2n}}, \tag{4.2.3}$$

where the notation

$$d_j = 2 \sum_{\ell=0}^{n-1} x_\ell \cos \frac{j(\ell + \frac{1}{2})\pi}{n} \qquad (4.2.4)$$

has been introduced to facilitate the discussion below.

Now, by the definition of the $2n$-point DFT in (4.1.8), we have

$$\left(\widehat{\widetilde{\mathbf{x}}} \right)_{2n-j} = \overline{\left(\widehat{\widetilde{\mathbf{x}}} \right)_j},$$

so that

$$\left(\widehat{\widetilde{\mathbf{x}}} \right)_{2n-j} = \overline{d_j e^{ij\frac{\pi}{2n}}} = d_j e^{-ij\frac{\pi}{2n}},$$

from which it follows by returning to (4.2.3), with j replaced by $2n - j$, that

$$\begin{aligned} d_{2n-j} &= e^{-i(2n-j)\frac{\pi}{2n}} \left(\widehat{\widetilde{\mathbf{x}}} \right)_{2n-j} \\ &= e^{-i(2n-j)\frac{\pi}{2n}} d_j e^{-ij\frac{\pi}{2n}} \\ &= e^{-i\pi} d_j = -d_j. \end{aligned} \qquad (4.2.5)$$

Thus, by setting $j = n$, we have

$$d_n = 0. \qquad (4.2.6)$$

Next, to recover $\widetilde{\mathbf{x}}$ from $\widehat{\widetilde{\mathbf{x}}}$ by taking the inverse $2n$-point DFT, as introduced also in Definition 1 on p.174 (see Theorem 1 on p.186), we have, by (4.2.3),

$$\begin{aligned} \left(\widetilde{\mathbf{x}} \right)_\ell &= \frac{1}{2n} \sum_{j=0}^{2n-1} \left(\widehat{\widetilde{\mathbf{x}}} \right)_j e^{i2\pi j\ell/2n} \\ &= \frac{1}{2n} \sum_{j=0}^{2n-1} d_j e^{ij\pi/2n} e^{ij\ell\pi/n} \\ &= \frac{1}{2n} \sum_{j=0}^{2n-1} d_j e^{ij(\ell+\frac{1}{2})\pi/n}. \end{aligned} \qquad (4.2.7)$$

Here, let us separate the summation (4.2.7) into two sums:

$$\sum_{j=0}^{2n-1} = \sum_{j=0}^{n-1} + \sum_{j=n}^{2n-1},$$

with

$$\sum_{j=n}^{2n-1} = \sum_{j=n}^{2n-1} d_j e^{ij(\ell+\frac{1}{2})\pi/n}$$

$$= \sum_{j=n+1}^{2n-1} d_j e^{ij(\ell+\frac{1}{2})\pi/n}$$

$$= \sum_{k=1}^{n-1} d_{2n-k} e^{i(2n-k)(\ell+\frac{1}{2})\pi/n},$$

where the term $d_n = 0$ in (4.2.6) reduces $\sum_{j=n}^{2n-1}$ to $\sum_{j=n+1}^{2n-1}$, and the change of summation index $k = 2n - j$ is applied to arrive at the final sum. Now observe that

$$e^{i(2n-k)(\ell+\frac{1}{2})\pi/n} = e^{i2(\ell+\frac{1}{2})\pi} e^{-ik(\ell+\frac{1}{2})\pi/n}$$

$$= -e^{-ik(\ell+\frac{1}{2})\pi/n}.$$

Hence, by (4.2.5) with j replaced by k, the expression in (4.2.7) becomes

$$\left(\widetilde{\mathbf{x}}\right)_\ell = \frac{1}{2n} d_0 + \frac{1}{2n} \sum_{j=1}^{n-1} d_j e^{ij(\ell+\frac{1}{2})\pi/n} + \frac{1}{2n} \sum_{k=1}^{n-1} (-d_k)(-e^{-ik(\ell+\frac{1}{2})\pi/n})$$

$$= \frac{1}{n} \left(\frac{d_0}{2} + \sum_{j=1}^{n-1} d_j \cos \frac{j(\ell+\frac{1}{2})\pi}{n} \right). \tag{4.2.8}$$

Therefore, since $x_\ell = \widetilde{x}_\ell = (\widetilde{\mathbf{x}})_\ell, \ell = 0, \ldots, n - 1$ by (4.2.1), we may return to the definition of d_j in (4.2.4) to re-formulate (4.2.8) as:

$$\sum_{k=0}^{n-1} x_k \delta_{k-\ell} = \frac{1}{n} \left\{ \sum_{k=0}^{n-1} x_k + 2 \sum_{j=1}^{n-1} \left(\sum_{k=0}^{n-1} x_k \cos \frac{j(k+\frac{1}{2})\pi}{n} \right) \cos \frac{j(\ell+\frac{1}{2})\pi}{2} \right\}$$

$$= \frac{1}{n} \sum_{k=0}^{n-1} \left\{ 1 + 2 \sum_{j=1}^{n-1} \cos \frac{j(k+\frac{1}{2})\pi}{n} \cos \frac{j(\ell+\frac{1}{2})\pi}{n} \right\} x_k$$

for all $\ell = 0, \ldots, n - 1$. Since $x_0, \ldots, x_{n-1}$ are arbitrarily chosen, it follows that

$$\frac{1}{n} \left\{ 1 + 2 \sum_{j=1}^{n-1} \cos \frac{j(k+\frac{1}{2})\pi}{n} \cos \frac{j(\ell+\frac{1}{2})\pi}{n} \right\} = \delta_{k-\ell}$$

for all $k, \ell = 0, \ldots, n - 1$. In other words, the n vectors $\mathbf{a}_0, \ldots, \mathbf{a}_{n-1} \in \mathbb{R}^n$, defined by

$$
\mathbf{a}_k = \left[\begin{array}{cccc} \dfrac{1}{\sqrt{n}} & \sqrt{\dfrac{2}{n}} \cos \dfrac{(k + \frac{1}{2})\pi}{n} & \cdots & \sqrt{\dfrac{2}{n}} \cos \dfrac{(n-1)(k + \frac{1}{2})\pi}{n} \end{array} \right]^T, \quad (4.2.9)
$$

where $k = 0, \ldots, n - 1$, constitute an orthonormal basis of $\mathbb{R}^n$; or equivalently, the $n \times n$ matrix

$$
C_n = [\mathbf{a}_0 \ \ldots \ \mathbf{a}_{n-1}], \quad (4.2.10)
$$

with column vectors $\mathbf{a}_0, \ldots, \mathbf{a}_{n-1}$, is an $n \times n$ orthogonal (or unitary) matrix (see Exercise 1).

Definition 1 $\boxed{\textbf{DCT \& IDCT}}$ *For each integer $n \geq 2$, the unitary matrix C_n, as introduced in (4.2.10), is called the n-point* **discrete cosine transformation** *(DCT), and its transpose C_n^T is called the n-point* **inverse discrete cosine transformation** *(IDCT).*

Since C_n is orthogonal (or unitary), we have

$$
C_n^T C_n = I_n,
$$

where I_n is the $n \times n$ identity matrix so that $C_n^{-1} = C_n^T$.

Example 1 Formulate and simplify the n-point DCT for $n = 2$ and 3.

Solution (i) For $n = 2$, we have, from (4.2.9),

$$
\mathbf{a}_k = \left[\begin{array}{cc} \dfrac{1}{\sqrt{2}} & \cos \dfrac{k + \frac{1}{2}}{2}\pi \end{array} \right]^T, \ k = 0, 1.
$$

Hence, the 2-point DCT is

$$
C_2 = \left[\begin{array}{cc} \dfrac{1}{\sqrt{2}} & \dfrac{1}{\sqrt{2}} \\ \cos \frac{\pi}{4} & \cos \frac{3\pi}{4} \end{array} \right] = \left[\begin{array}{cc} \dfrac{1}{\sqrt{2}} & \dfrac{1}{\sqrt{2}} \\ \dfrac{1}{\sqrt{2}} & \dfrac{-1}{\sqrt{2}} \end{array} \right] = \dfrac{1}{\sqrt{2}} \left[\begin{array}{cc} 1 & 1 \\ 1 & -1 \end{array} \right]
$$

with IDCT $C_2^{-1} = C_2^T$. Observe that

$$
C_2^T C_2 = \left[\begin{array}{cc} \dfrac{1}{\sqrt{2}} & \dfrac{1}{\sqrt{2}} \\ \dfrac{1}{\sqrt{2}} & \dfrac{-1}{\sqrt{2}} \end{array} \right] \left[\begin{array}{cc} \dfrac{1}{\sqrt{2}} & \dfrac{1}{\sqrt{2}} \\ \dfrac{1}{\sqrt{2}} & \dfrac{-1}{\sqrt{2}} \end{array} \right] = \left[\begin{array}{cc} 1 & 0 \\ 0 & 1 \end{array} \right] = I_2.
$$

(ii) For $n = 3$, we have, from (4.2.9),

$$
\mathbf{a}_k = \left[\begin{array}{ccc} \dfrac{1}{\sqrt{3}} & \sqrt{\dfrac{2}{3}} \cos \dfrac{k + \frac{1}{2}}{3}\pi & \sqrt{\dfrac{2}{3}} \cos \dfrac{2(k + \frac{1}{2})}{3}\pi \end{array} \right]^T,
$$

for $k = 0, 1, 2$. Hence, the 3-point DCT is

$$
C_3 = \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \sqrt{\frac{2}{3}}\cos\frac{\pi}{6} & \sqrt{\frac{2}{3}}\cos\frac{3\pi}{6} & \sqrt{\frac{2}{3}}\cos\frac{5\pi}{6} \\ \sqrt{\frac{2}{3}}\cos\frac{2\pi}{6} & \sqrt{\frac{2}{3}}\cos\frac{6\pi}{6} & \sqrt{\frac{2}{3}}\cos\frac{10\pi}{6} \end{bmatrix}
$$

$$
= \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{6}} & -\sqrt{\frac{2}{3}} & \frac{1}{\sqrt{6}} \end{bmatrix}.
$$

Observe that

$$
C_3^T C_3 = \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & 0 & -\sqrt{\frac{2}{3}} \\ \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{6}} & -\sqrt{\frac{2}{3}} & \frac{1}{\sqrt{6}} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \qquad \blacksquare
$$

Example 2 Compute the DCT of the vector $\mathbf{x} = [1 \ -3 \ 3]^T \in \mathbb{R}^3$.

Solution   By the solution of Example 1, we have

$$
\widehat{\mathbf{x}} = C_3\mathbf{x} = \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{6}} & -\sqrt{\frac{2}{3}} & \frac{1}{\sqrt{6}} \end{bmatrix} \begin{bmatrix} 1 \\ -3 \\ 3 \end{bmatrix}
$$

$$
= \begin{bmatrix} \frac{1}{\sqrt{3}}(1 - 3 + 3) \\ \frac{1}{\sqrt{2}}(1 + 0 - 3) \\ \frac{1}{\sqrt{6}}(1 + 3) + 3\sqrt{\frac{2}{3}} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{3}} \\ -\sqrt{2} \\ \frac{10}{\sqrt{6}} \end{bmatrix}.
$$

Observe that the IDCT of $\widehat{\mathbf{x}}$ is
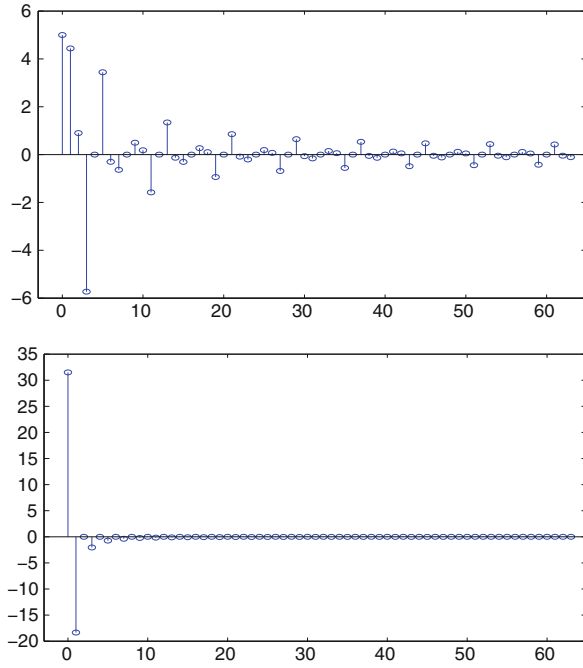
Fig. 4.3 *Top* DCT of S_1; *Bottom* DCT of S_2

$$
C_3^T \widehat{\mathbf{x}} =
\begin{bmatrix}
\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \\
\frac{1}{\sqrt{3}} & 0 & -\sqrt{\frac{2}{3}} \\
\frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}}
\end{bmatrix}
\begin{bmatrix}
\frac{1}{\sqrt{3}} \\
-\sqrt{2} \\
\frac{10}{\sqrt{6}}
\end{bmatrix}
$$

$$
=
\begin{bmatrix}
\frac{1}{3} - 1 + \frac{10}{6} \\
\frac{1}{3} - \frac{\sqrt{2}}{3} \cdot \frac{10}{\sqrt{2}\sqrt{3}} \\
\frac{1}{3} + 1 + \frac{10}{6}
\end{bmatrix}
=
\begin{bmatrix}
1 \\
-3 \\
3
\end{bmatrix}
= \mathbf{x}.
$$

∎

Example 3 Let S_1, $S_2 \in \mathbb{R}^{64}$ be signals considered in Example 3 on p.177. The DCTs of S_1 and S_2 are shown in Fig. 4.3. ∎

Remark 1 Let $\mathbf{x} \in \mathbb{R}^n$. Since the n-point DCT of $\mathbf{x}$, defined by

$$\widehat{\mathbf{x}} = C_n \mathbf{x},$$

is computed by taking the inner products of the row vectors of C_n with $\mathbf{x}$, it is more convenient to re-formulate the DCT (orthogonal, i.e. unitary matrix), defined in (4.2.10), in terms of row vectors; namely,

$$C_n = [\mathbf{a}_0 \ \cdots \ \mathbf{a}_{n-1}] = \begin{bmatrix} \mathbf{c}_0^T \\ \vdots \\ \mathbf{c}_{n-1}^T \end{bmatrix}, \tag{4.2.11}$$

where $\mathbf{c}_0^T$ is the first row of C_n, $\mathbf{c}_1^T$ is the second row of C_n, ..., and $\mathbf{c}_{n-1}^T$ is the nth row of C_n. More precisely, $\mathbf{c}_j^T$ is the transpose of the column vector $\mathbf{c}_j$ defined as follows:

$$\mathbf{c}_0 = \left[\frac{1}{\sqrt{n}} \ \cdots \ \frac{1}{\sqrt{n}} \right]^T = \frac{1}{\sqrt{n}} [1 \ \dots \ 1]^T, \tag{4.2.12}$$

and for all $j = 1, \dots, n-1$,

$$\mathbf{c}_j = \sqrt{\frac{2}{n}} \left[\cos \frac{j\pi}{2n} \ \cos \frac{j3\pi}{2n} \ \cdots \ \cos \frac{j(2n-1)\pi}{2n} \right]^T. \tag{4.2.13}$$

Then it follows from (4.2.10) and (4.2.11) that the corresponding IDCT can be re-formulated as

$$C_n^{-1} = C_n^T = [\mathbf{c}_0 \ \dots \ \mathbf{c}_{n-1}] \tag{4.2.14}$$

with column vectors $\mathbf{c}_0, \cdots, \mathbf{c}_{n-1}$ introduced in (4.2.12) and (4.2.13).   ∎

Theorem 1 *Let* $\mathbf{x} \in \mathbb{R}^n$. *The* n-point *DCT of* $\mathbf{x}$ *is given by*

$$\widehat{\mathbf{x}} = C_n \mathbf{x} = \begin{bmatrix} \langle \mathbf{x}, \mathbf{c}_0 \rangle \\ \vdots \\ \langle \mathbf{x}, \mathbf{c}_{n-1} \rangle \end{bmatrix}, \tag{4.2.15}$$

and $\mathbf{x}$ *can be written as*

$$\mathbf{x} = \sum_{j=0}^{n-1} \langle \mathbf{x}, \mathbf{c}_j \rangle \mathbf{c}_j. \tag{4.2.16}$$

Indeed, (4.2.15) follows from matrix multiplication by using the formulation of C_n in (4.2.11), and (4.2.16) is a re-formulation of

$$\mathbf{x} = C_n^T (C_n \mathbf{x}) = [\mathbf{c}_0 \ \cdots \ \mathbf{c}_{n-1}] \begin{bmatrix} \langle \mathbf{x}, \mathbf{c}_0 \rangle \\ \vdots \\ \langle \mathbf{x}, \mathbf{c}_{n-1} \rangle \end{bmatrix},$$

which is obtained by applying (4.2.14) and (4.2.15).   ∎

Definition 2 ⎹ **DCT as inner product** ⎸ *For any* $\mathbf{x} \in \mathbb{R}^n$, *where* $n \geq 2$ *is an arbitrary* (*fixed*) *integer,* $\widehat{\mathbf{x}}$ *as defined by* (4.2.15) *is called the* n-point *DCT of* $\mathbf{x}$ *and the* (*finite*) *series expansion in* (4.2.16) *is called the discrete cosine series representation*

of **x***, with cosine coefficients* $\langle \mathbf{c}_0, \mathbf{x} \rangle, \ldots, \langle \mathbf{c}_{n-1}, \mathbf{x} \rangle$*. Furthermore, the first cosine coefficient* $\langle \mathbf{c}_0, \mathbf{x} \rangle$ *is called the* **DC** (**direct current**) **term** *and the remaining cosine coefficients are called the* **AC** (**alternate current**) **terms** *of* $\widehat{\mathbf{x}}$*.*

Remark 2 The DC term of $\widehat{\mathbf{x}}$ for any given $\mathbf{x} = (x_0, \ldots, x_{n-1}) \in \mathbb{R}^n$ is given by

$$\langle \mathbf{x}, \mathbf{c}_0 \rangle = \frac{1}{\sqrt{n}} \sum_{k=0}^{n-1} x_k,$$

which is simply the "average" of the components of $\mathbf{x}$, while the AC terms are governed by some non-trivial (oscillating) cosines (see Remarks 3–4 to follow). ∎

Remark 3 As a continuation of Remark 2, consider the n-point DCT of a vector $\mathbf{f} \in \mathbb{R}^n$ obtained by "sampling" a continuous function $f(t)$ for $0 \le t \le 1$, at the sample points

$$t_k = \frac{2k+1}{2n}, \ k = 0, \ldots, n-1; \tag{4.2.17}$$

namely, $\mathbf{f} = (f(t_0), \ldots, f(t_{n-1})) \in \mathbb{R}^n$. Then by applying the finite cosine series representation (4.2.16), we have

$$f(t_\ell) = \frac{b_0}{2} + \sum_{j=1}^{n-1} b_j \cos(j\pi t_\ell), \ \text{for all } \ell = 0, \ldots, n-1, \tag{4.2.18}$$

where

$$b_j = \frac{2}{n} \sum_{k=0}^{n-1} f(t_k) \cos(j\pi t_k), \ j = 0, \ldots, n-1 \tag{4.2.19}$$

(see Exercise 8). ∎

Remark 4 Since $\cos 2\pi t$ is periodic with period $= 1$, the frequency of $\cos(j\pi t)$ is $\frac{1}{2}j$ Hz (see (4.1.1)). Hence, when $f(t)$ in Remark 3 is considered as a traveling wave, then the finite cosine series representation of $f(t)$ (at the time samples $t = t_\ell$ as in (4.2.17)) reveals the various frequency components of $f(t)$: with stationary (or DC) component $b_0/2$ and traveling (or AC) components $b_1, b_2, \ldots, b_{n-1}$ at frequencies $\frac{1}{2}, \frac{2}{2}, \ldots, \frac{n-1}{2}$, respectively. ∎

Remark 5 The continuous (or analog) version of the finite cosine series representation in (4.2.18), obtained by replacing t_ℓ in (4.2.18) with $t \in [0, 1]$ is called the (Fourier) cosine series of $f \in C[0, 1]$; namely,

$$f(t) = \frac{b_0}{2} + \sum_{j=1}^{\infty} b_j \cos j\pi t.$$

In addition, when the sum in (4.2.19) is replaced by the integral over [0, 1], the coefficients $b_0, b_1, \ldots$ (of the above cosine series) are called (Fourier) cosine coefficients. That is, the continuous version of (4.2.19) is given by

$$b_j = 2 \int_0^1 f(t) \cos j\pi t \, dt.$$

In other words, the n-point DCT of an n-vector $\mathbf{x}$ can be considered as discretization of the cosine coefficients of $x(t)$ at $t_0, \ldots, t_{n-1}$ given by (4.2.17), where $\mathbf{x} = [x(t_0), \ldots, x(t_{n-1})]^T$. The topic of Fourier cosine series will be studied in Sect. 6.2.   ∎

<div align="center">

Exercises

</div>

Exercise 1 Apply the n-point DCT C_n, for $n = 2$ and 3, in Example 1 to compute the discrete cosine transforms of the following data sets.

(a) $\widehat{\mathbf{x}}_j = C_2 \mathbf{x}_j$ for $j = 1, 2, 3$, where

$$\mathbf{x}_1 = [\,1\ 2\,]^T, \quad \mathbf{x}_2 = [\,2\ 1\,]^T, \quad \mathbf{x}_3 = [\,-1\ 1\,]^T.$$

(b) $\widehat{\mathbf{y}}_j = C_3 \mathbf{y}_j$ for $j = 1$ and 2, where

$$\mathbf{y}_1 = [\,1\ 2\ 3\,]^T, \quad \mathbf{y}_2 = [\,-1\ 2\ 1\,]^T.$$

Exercise 2 Apply the n-point IDCT $C_n^{-1} = C_n^T$, as introduced in Definition 1, to verify your answers in Exercise 1, by computing $C_2^{-1} \mathbf{x}_j$ for $j = 1, 2, 3$ in (a) above, and $C_3^{-1} \mathbf{y}_j$ for $j = 1, 2$ in (b) above.

Exercise 3 Let $\widehat{\mathbf{x}} = C_n \mathbf{x}$ be the n-point DCT of $\mathbf{x} \in \mathbb{R}^n$, where

$$C_n = [\mathbf{a}_0 \ \cdots \ \mathbf{a}_{n-1}]$$

with the n column vectors $\mathbf{a}_0, \ldots, \mathbf{a}_{n-1}$ defined as in (4.2.9). Let $\mathbf{c}_0^T, \ldots, \mathbf{c}_{n-1}^T$ denote the row vectors of C_n; that is,

$$C_n = \begin{bmatrix} \mathbf{c}_0^T \\ \vdots \\ \mathbf{c}_{n-1}^T \end{bmatrix}$$

as in (4.2.11). Verify that the vectors $\mathbf{c}_0, \ldots, \mathbf{c}_{n-1}$ are given by (4.2.12)–(4.2.13) and illustrate this result by writing out such rows for the cases $n = 2$ and 3.

Exercise 4 As a continuation of Exercise 3, compute the following DCT, namely:

(a) $\widehat{\mathbf{x}}_j = C_2 \mathbf{x}_j$, $j = 1, 2, 3$, for $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ given in Exercise 1, by taking inner products with $\mathbf{c}_0, \mathbf{c}_1, \mathbf{c}_2$ as in Theorem 1; that is,

$$\widehat{\mathbf{x}}_j = [\langle \mathbf{c}_0, \mathbf{x}_j \rangle \ \langle \mathbf{c}_1, \mathbf{x}_j \rangle \ \langle \mathbf{c}_2, \mathbf{x}_j \rangle]^T;$$

(b) $\widehat{\mathbf{y}}_j = C_3 \mathbf{y}_j$, $j = 1, 2$ for $\mathbf{y}_1, \mathbf{y}_2$ given in Exercise 1, by taking inner products with the vectors $\mathbf{c}_j$'s instead of the $\mathbf{a}_j$'s; that is,

$$\widehat{\mathbf{y}}_j = [\langle \mathbf{c}_0, \mathbf{y}_j \rangle \ \langle \mathbf{c}_1, \mathbf{y}_j \rangle]^T.$$

Exercise 5 As a continuation of Exercise 4, compute the IDCT of $\widehat{\mathbf{x}}_j$ and $\widehat{\mathbf{y}}_j$ simply by writing out and adding the following sums:

(a) $\mathbf{x}_j = \langle \mathbf{c}_0, \mathbf{x}_j \rangle \mathbf{c}_0 + \langle \mathbf{c}_1, \mathbf{x}_j \rangle \mathbf{c}_1 + \langle \mathbf{c}_2, \mathbf{x}_j \rangle \mathbf{c}_2$ for $j = 1, 2, 3$; and
(b) $\mathbf{y}_j = \langle \mathbf{c}_0, \mathbf{y}_j \rangle \mathbf{c}_0 + \langle \mathbf{c}_1, \mathbf{y}_j \rangle \mathbf{c}_1$ for $j = 1, 2$.
(c) Compare the results in (a) and (b) with those in Exercise 2. Note that the vectors $\mathbf{c}_j$'s in (a) are different from those in (b), since the ones in (a) are for 2-point DCT, while those in (b) are for 3-point DCT.

Exercise 6 For each of the following continuous functions $f(x)$ on $[0, 1]$, compute its discrete cosine coefficients

$$b_j = \frac{2}{3} \sum_{k=0}^{2} f\left(\frac{2k+1}{6}\right) \cos\left(\frac{j(2k+1)\pi}{6}\right)$$

for $j = 0, 1, 2$. Then compute the finite cosine Fourier series of f, namely:

$$\tilde{f}\left(\frac{2\ell+1}{6}\right) = \frac{b_0}{2} + \sum_{j=0}^{2} b_j \cos\left(\frac{j(2\ell+1)\pi}{6}\right)$$

for $\ell = 0, 1, 2$.

(a) $f(x) = |x - \frac{1}{2}|$.
(b) $f(x) = 2x^2 - 2x + 1$.

Exercise 7 As a continuation of Exercise 6, compare the values of $\tilde{f}\left(\frac{2\ell+1}{6}\right)$ with $f\left(\frac{2\ell+1}{6}\right)$, $\ell = 0, 1, 2$, for the two functions $f(x) = |x - \frac{1}{2}|$ and $f(x) = 2x^2 - 2x + 1$. (Note that only the special case $n = 3$ is considered in this example. For large values of n, the finite cosine Fourier series introduced in (4.2.18) of Remark 3 provide good approximation of these two functions. See Exercise 8 below.)

Exercise 8 Use MATLAB to compute the discrete cosine coefficients

$$b_j = \frac{2}{n} \sum_{k=0}^{n-1} f\left(\frac{2k+1}{2n}\right) \cos\left(\frac{j(2k+1)\pi}{2n}\right)$$

for the functions $f(x) = |x - \frac{1}{2}|$, for $n \geq 100$. Then compute the values of the finite cosine Fourier series

$$\tilde{f}\left(\frac{2\ell + 1}{2n}\right) = \frac{b_0}{2} + \sum_{j=0}^{n-1} b_j \cos\left(\frac{j(2\ell + 1)\pi}{2n}\right)$$

with the corresponding values

$$f\left(\frac{2\ell + 1}{2n}\right) = \left|\frac{2\ell + 1}{2n} - \frac{1}{2}\right|$$

for $\ell = 0, \ldots, n - 1$.

Exercise 9 Repeat Exercise 8 for the function $f(x) = 2x^2 - 2x + 1$ in Exercise 6 (b).

4.3 Fast Fourier Transform

To study the FFT computational scheme, we need the following result due to Cornelius Lanczos to reduce a $2n$-point DFT to an n-point DFT via multiplication by certain sparse matrices.

Denote
$$D_n = \text{diag}\{1, e^{-i\pi/n}, \ldots, e^{-i(n-1)\pi/n}\},$$

and let

$$P_n^e = [\delta_{2j-k}] = \begin{bmatrix} 1\,0\,0\,0 & \cdots & 0 \\ 0\,0\,1\,0 & \cdots & 0 \\ & \cdots & \\ 0\,0\,0\ldots & 1 & 0 \end{bmatrix};$$

$$P_n^o = [\delta_{2j-k+1}] = \begin{bmatrix} 0\,1\,0\,0 & \cdots & 0 \\ 0\,0\,0\,1 & \cdots & 0 \\ & \cdots & \\ 0\,0 & \cdots\,0 & 1 \end{bmatrix},$$

where $0 \leq j \leq n - 1$ and $0 \leq k \leq 2n - 1$, be $n \times (2n)$ matrices. Also, let I_n denote the $n \times n$ identity matrix and

$$E_{2n} = \begin{bmatrix} I_n & \vdots & D_n \\ \cdots\cdots & \cdots\cdots \\ I_n & \vdots & -D_n \end{bmatrix}. \qquad (4.3.1)$$

For example, with the above notations, we have

$$D_1 = [1], \ P_1^e = [1 \ 0], \ P_1^o = [0 \ 1], \ E_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix},$$

$$D_2 = \begin{bmatrix} 1 & 0 \\ 0 & -i \end{bmatrix}, \ P_2^e = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \ P_2^o = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$E_4 = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & -i \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & i \end{bmatrix}.$$

Theorem 1 **Factorization of DFT matrices** *The 2n-point DFT F_{2n} can be factored out in terms of two diagonal blocks of the n-point DFT F_n as follows:*

$$F_{2n} = E_{2n} \begin{bmatrix} F_n & \vdots & O \\ \cdots\cdots\cdots\cdots \\ O & \vdots & F_n \end{bmatrix} \begin{bmatrix} P_n^e \\ \cdots \\ P_n^o \end{bmatrix}, \tag{4.3.2}$$

where O denotes the $n \times n$ zero matrix.

Proof To derive (4.3.2), consider $\mathbf{x} = [x_0, \ \ldots, \ x_{2n-1}]^T \in \mathbb{C}^{2n}$, with DFT given by $\widehat{\mathbf{x}} = F_{2n}\mathbf{x} = [\widehat{x}_0, \ \ldots, \ \widehat{x}_{2n-1}]^T$, where

$$\begin{aligned}
\widehat{x}_\ell &= \sum_{k=0}^{2n-1} x_k \, e^{-i2\pi k\ell/(2n)} \\
&= \sum_{j=0}^{n-1} x_{2j} e^{-i2\pi(2j)\ell/(2n)} + \sum_{j=0}^{n-1} x_{2j+1} \, e^{-i2\pi(2j+1)\ell/(2n)} \\
&= \sum_{j=0}^{n-1} x_{2j} e^{-i2\pi j\ell/n} + e^{-i\pi\ell/n} \sum_{j=0}^{n-1} x_{2j+1} \, e^{-i2\pi j\ell/n}.
\end{aligned} \tag{4.3.3}$$

Set

$$\mathbf{x}_e = [x_0, \ x_2, \ \ldots, \ x_{2n-2}]^T, \ \mathbf{x}_o = [x_1, \ x_3, \ \ldots, \ x_{2n-1}]^T.$$

Then

$$\mathbf{x}_e = P_n^e \mathbf{x}, \ \mathbf{x}_o = P_n^o \mathbf{x}.$$

Using the notation $(\mathbf{v})_\ell$ for the ℓth component of the vector $\mathbf{v}$, the result in (4.3.3) is simply

$$
\begin{aligned}
\widehat{x}_\ell &= (F_n \mathbf{x}_e)_\ell + e^{-i\pi\ell/n}(F_n \mathbf{x}_o)_\ell \\
&= (F_n P_n^e \mathbf{x})_\ell + e^{-i\pi\ell/n}(F_n P_n^o \mathbf{x})_\ell \\
&= \begin{cases} (F_n P_n^e \mathbf{x})_\ell + e^{-i\pi\ell/n}(F_n P_n^o \mathbf{x})_\ell, & \text{for } 0 \le \ell \le n-1, \\ (F_n P_n^e \mathbf{x})_\ell - e^{-i\pi(\ell-n)/n}(F_n P_n^o \mathbf{x})_\ell, & \text{for } n \le \ell \le 2n-1. \end{cases}
\end{aligned} \tag{4.3.4}
$$

In (4.3.4), we have used the fact that $e^{-i\pi\ell/n} = -e^{-i\pi(\ell-n)/n}$. Thus,

$$
(F_{2n}\mathbf{x})_\ell = \begin{cases} \left((F_n P_n^e + D_n F_n P_n^o)\mathbf{x}\right)_\ell, & \text{for } 0 \le \ell \le n-1, \\ \left((F_n P_n^e - D_n F_n P_n^o)\mathbf{x}\right)_\ell, & \text{for } n \le \ell \le 2n-1, \end{cases}
$$

which yields (4.3.2).   ∎

Let $n = 2^m$ with $m \ge 1$. For each k, $0 \le k \le m-1$, let G_k^m be the $2^m \times 2^m$ matrix defined by

$$
G_k^m = \text{diag}\{\underbrace{E_{2^{m-k}}, \ \ldots, \ E_{2^{m-k}}}_{2^k \text{ copies}}\} = \begin{bmatrix} E_{2^{m-k}} & & O \\ & \ddots & \\ O & & E_{2^{m-k}} \end{bmatrix}, \tag{4.3.5}
$$

where, according to (4.3.1), $E_{2^{m-k}}$ is a $2^{m-k} \times 2^{m-k}$ matrix:

$$
E_{2^{m-k}} = \begin{bmatrix} I_{2^{m-k-1}} & \vdots & D_{2^{m-k-1}} \\ \cdots\cdots & \cdots & \cdots \\ I_{2^{m-k-1}} & \vdots & -D_{2^{m-k-1}} \end{bmatrix}.
$$

Furthermore, let

$$
P_{2^m} = \begin{bmatrix} P_{2^{m-1}}^e \\ \cdots \\ P_{2^{m-1}}^o \end{bmatrix}
$$

denote the $2^m \times 2^m$ "permutation matrix" in (4.3.2) with $n = 2^m$; and define, inductively, the permutation matrices $\widetilde{P}_1 = \widetilde{P}_{2^0}, \widetilde{P}_2 = \widetilde{P}_{2^1}, \widetilde{P}_4 = \widetilde{P}_{2^2}, \ldots, \widetilde{P}_{2^m}$ by

$$
\widetilde{P}_1 = [1], \ldots, \widetilde{P}_{2^\ell} = \begin{bmatrix} \widetilde{P}_{2^{\ell-1}} & O \\ O & \widetilde{P}_{2^{\ell-1}} \end{bmatrix} P_{2^\ell},
$$

where $\ell = 1, 2, \ldots, m$.

Example 1   Consider $n = 4$, $m = 2$. We have

$$G_0^2 = E_{2^2} = E_4,$$

$$G_1^2 = \text{diag}\{E_2, E_2\} = \begin{bmatrix} E_2 & O \\ O & E_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & -1 \end{bmatrix},$$

and

$$P_2 = \begin{bmatrix} P_1^e \\ P_1^o \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

$$\widetilde{P}_2 = \begin{bmatrix} \widetilde{P}_1 & 0 \\ 0 & \widetilde{P}_1 \end{bmatrix} P_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} P_2 = P_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

$$P_4 = \begin{bmatrix} P_2^e \\ P_2^o \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$\widetilde{P}_4 = \begin{bmatrix} \widetilde{P}_2 & O \\ O & \widetilde{P}_2 \end{bmatrix} P_4 = \begin{bmatrix} I_2 & O \\ O & I_2 \end{bmatrix} P_4 = P_4. \qquad \blacksquare$$

Next we state the fast Fourier transform (FFT) computational scheme as follows.

Theorem 2 ⎢**Full factorization of DFT matrices**⎥ *Let $n = 2^m$, where $m \geq 1$ is an integer. Then the n-point DFT has the formulation*

$$F_n = F_{2^m} = G_0^m G_1^m \dots G_{m-1}^m \widetilde{P}_{2^m}. \tag{4.3.6}$$

Proof   Proof of the FFT scheme (4.3.6) can be carried out by mathematical induction on $m = 1, 2, \dots$.

For $m = 1$, since

$$G_0^1 = E_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

and $\widetilde{P}_2 = I_2$ as shown in Example 1, it follows that

$$G_0^1 \widetilde{P}_2 = G_0^1 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} = F_2,$$

which is (4.3.6) for $m = 1$.

In general, by Theorem 1 and the induction hypothesis, we have

$$
\begin{aligned}
F_{2^m} &= E_{2^m} \begin{bmatrix} F_{2^{m-1}} & O \\ O & F_{2^{m-1}} \end{bmatrix} P_{2^m} \\
&= E_{2^m} \begin{bmatrix} G_0^{m-1} \dots G_{m-2}^{m-1} \widetilde{P}_{2^{m-1}} & O \\ O & G_0^{m-1} \dots G_{m-2}^{m-1} \widetilde{P}_{2^{m-1}} \end{bmatrix} P_{2^m} \\
&= E_{2^m} \begin{bmatrix} G_0^{m-1} & O \\ O & G_0^{m-1} \end{bmatrix} \dots \begin{bmatrix} G_{m-2}^{m-1} & O \\ O & G_{m-2}^{m-1} \end{bmatrix} \begin{bmatrix} \widetilde{P}_{2^{m-1}} & O \\ O & \widetilde{P}_{2^{m-1}} \end{bmatrix} P_{2^m} \\
&= E_{2^m} G_1^m G_2^m \dots G_{m-1}^m \widetilde{P}_{2^m} \\
&= G_0^m G_1^m \dots G_{m-1}^m \widetilde{P}_{2^m},
\end{aligned}
$$

since $G_0^m = E_{2^m}$ by (4.3.5), and

$$
\begin{bmatrix} G_{k-1}^{m-1} & O \\ O & G_{k-1}^{m-1} \end{bmatrix} = G_k^m, \tag{4.3.7}
$$

for $1 \le k \le m-1$ (see Exercise 2).   ∎

Remark 1 Since the IDFT is given by $\widetilde{F}_n = \frac{1}{n} F_n^*$, it follows from (4.3.6) that for $n = 2^m$,

$$
\widetilde{F}_n = \widetilde{F}_{2^m} = 2^{-m} \widetilde{P}_{2^m}^T (G_{m-1}^m)^* \dots (G_0^m)^*. \tag{4.3.8}
$$

Example 2 Let us verify (4.3.6) for $n = 4, n = 8$. First, by Theorem 1,

$$
F_4 = E_4 \begin{bmatrix} F_2 & O \\ O & F_2 \end{bmatrix} \begin{bmatrix} P_2^e \\ P_2^o \end{bmatrix} = E_4 \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & -1 \end{bmatrix} P_4 = G_0^2 G_1^2 \widetilde{P}_4,
$$

where the last equality follows from $G_0^2 = E_4$, the expression of G_1^2 and $P_4 = \widetilde{P}_4$ as shown in Example 1. This is (4.3.6) for $n = 4$.

For $n = 8$,

$$
\begin{aligned}
F_8 &= E_8 \begin{bmatrix} F_4 & O \\ O & F_4 \end{bmatrix} \begin{bmatrix} P_4^e \\ P_4^o \end{bmatrix} = E_8 \begin{bmatrix} G_0^2 G_1^2 \widetilde{P}_4 & O \\ O & G_0^2 G_1^2 \widetilde{P}_4 \end{bmatrix} P_8 \\
&= E_8 \begin{bmatrix} G_0^2 & O \\ O & G_0^2 \end{bmatrix} \begin{bmatrix} G_1^2 & O \\ O & G_1^2 \end{bmatrix} \begin{bmatrix} \widetilde{P}_4 & O \\ O & \widetilde{P}_4 \end{bmatrix} P_8 \\
&= G_0^3 G_1^3 G_2^3 \widetilde{P}_8,
\end{aligned}
$$

where the last equality follows from (4.3.7).   ∎

Remark 2 | **Complexity of FFT** | Computational complexity is the operation count, including the number of long operations (i.e. multiplication and division) and short operations (i.e. addition and subtraction). To compute the n-point DFT, for $F_n = \left[(z_n)^{jk}\right]$, if we count $(z_n)^{jk}c$ for a $c \in \mathbb{C}$ as one (complex) multiplication operation even if $(z_n)^{jk} = 1$ (when $jk = 0$ or $jk = n\ell$ for some integer $\ell > 0$), then the number of multiplication operations in computing an n-point DFT is n^2. The significance of the special case of $n = 2^m$ is that the FFT computational scheme, as formulated in (4.3.6) of Theorem 2, can be applied to compute the n-point DFT with only $O(mn) = O(n \log n)$ operations (see Exercise 1). The count in terms of the order of magnitudes $O(n^2)$ and $O(n \log n)$ is valid only for large values of n. In many applications, a signal is partitioned into segments, each of which is of length n, where n is relatively small. Hence, a precise count is more valuable, particularly for hardware implementation. In the following, we give a precise count and implementation-ready algorithms for $n = 4, 8, 16$ for computing the n-point DFT via the FFT scheme. The 16-point DFT, F_{16}, is particularly important, since fast computation of the 8-point DCT depends on the computation of F_{16}. The topic of fast discrete cosine transform (FDCT) will be studied in the next section. Application of 8-point DCT to image and video compressions will be the topic of discussion in Sect. 5.4. ∎

| **Operation count for 4-point FFT** | For $\mathbf{x} = [x_0, x_1, x_2, x_3]^T \in \mathbb{C}^4$, let $\mathbf{g} = \widehat{\mathbf{x}}$ denote its DFT:

$$\mathbf{g} = F_4\mathbf{x} = E_4 \operatorname{diag}\{E_2, E_2\}\, P_4\mathbf{x}.$$

The role of P_4 is to interchange the orders of x_j in $\mathbf{x}$, namely:

$$P_4\mathbf{x} = \begin{bmatrix} x_0 \\ x_2 \\ x_1 \\ x_3 \end{bmatrix}.$$

Denote $\mathbf{d} = \operatorname{diag}\{E_2, E_2\}\, P_4\mathbf{x}$. Then $\mathbf{g} = E_4\mathbf{d}$, and the FFT algorithm to compute $\mathbf{g} = \widehat{\mathbf{x}} = F_4\mathbf{x}$ is

$$
\begin{cases}
d_0 = x_0 + x_2 \\
d_1 = x_0 - x_2 \\
d_2 = x_1 + x_3 \\
d_3 = x_1 - x_3
\end{cases}
\qquad
\begin{cases}
g_0 = d_0 + d_2 \\
g_1 = d_1 - id_3 \\
g_2 = d_0 - d_2 \\
g_3 = d_1 + id_3.
\end{cases}
\tag{4.3.9}
$$

Observe that we need one complex multiplication id_3 in (4.3.9) for computing g_1; but complex multiplication is not required to compute g_3 since id_3 has already been

computed. Thus, the FFT for $n = 4$ requires a total of one complex multiplication and 8 additions.   ■

Operation count for 8-point FFT For $n = 8$, denote $W = z_8 = e^{-i\frac{2\pi}{8}} = \frac{\sqrt{2}}{2} - i\frac{\sqrt{2}}{2}$. From Theorem 2 (see also Example 2), we have

$$F_8 = G_0^3 G_1^3 G_2^3 \widetilde{P}_8,$$

where

$$G_0^3 = E_8 = \begin{bmatrix} I_4 & D_4 \\ I_4 & -D_4 \end{bmatrix} \text{ with } D_4 = \text{diag}\{1, W, -i, -\overline{W}\},$$

$$G_1^3 = \text{diag}\{G_0^2, G_0^2\} \text{ with } G_0^2 = E_4 = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & -i \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & i \end{bmatrix},$$

$$G_2^3 = \text{diag}\{G_1^2, G_1^2\} \text{ with } G_1^2 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & -1 \end{bmatrix},$$

and

$$\widetilde{P}_8 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \tag{4.3.10}$$

The role of $\widetilde{P}_8$ is to interchange the orders of x_j in $\mathbf{x} = [x_0, x_1, \ldots, x_7]^T \in \mathbb{C}^8$ such that

$$\widetilde{P}_8 \mathbf{x} = [x_0, x_4, x_2, x_6, x_1, x_5, x_3, x_7]^T.$$

Let $\mathbf{d}, \mathbf{h}, \mathbf{g} \in \mathbb{C}^8$ denote

$$\mathbf{d} = G_2^3 \widetilde{P}_8 \mathbf{x}, \ \ \mathbf{h} = G_1^3 \mathbf{d}, \ \ \mathbf{g}(= \widehat{\mathbf{x}}) = G_0^3 \mathbf{h}.$$

Then, with $W = \frac{\sqrt{2}}{2} - i\frac{\sqrt{2}}{2}$, the FFT algorithm for computing the 8-point DFT $\mathbf{g} = \widehat{\mathbf{x}} = F_8 \mathbf{x}$ can be realized as follows:

$$
\begin{cases} d_0 = x_0 + x_4 \\ d_1 = x_0 - x_4 \\ d_2 = x_2 + x_6 \\ d_3 = x_2 - x_6 \\ d_4 = x_1 + x_5 \\ d_5 = x_1 - x_5 \\ d_6 = x_3 + x_7 \\ d_7 = x_3 - x_7 \end{cases}
\begin{cases} h_0 = d_0 + d_2 \\ h_1 = d_1 - i d_3 \\ h_2 = d_0 - d_2 \\ h_3 = d_1 + i d_3 \\ h_4 = d_4 + d_6 \\ h_5 = d_5 - i d_7 \\ h_6 = d_4 - d_6 \\ h_7 = d_5 + i d_7 \end{cases}
\begin{cases} g_0 = h_0 + h_4 \\ g_1 = h_1 + W h_5 \\ g_2 = h_2 - i h_6 \\ g_3 = h_3 - \overline{W} h_7 \\ g_4 = h_0 - h_4 \\ g_5 = h_1 - W h_5 \\ g_6 = h_2 + i h_6 \\ g_7 = h_3 + \overline{W} h_7. \end{cases}
\tag{4.3.11}
$$

Observe that we need five complex multiplications $i d_3$, $i d_7$, $W h_5$, $i h_6$ and $\overline{W} h_7$ in (4.3.11) for computing h_1, h_5, g_1, g_2, g_3, and that the values of $i d_3$, $i d_7$, $W h_5$, $i h_6$ and $\overline{W} h_7$ are stored in the memory RAM for computing h_3, h_7, g_5, g_6, g_7. Thus, the computational complexity for computing the 8-point DFT via FFT consists of 5 complex multiplications and 24 additions. ∎

Operation count for 16-point FFT For $n = 16$, denote $w = z_{16} = e^{-i \frac{\pi}{8}}$. Thus, $W = \frac{\sqrt{2}}{2} - i \frac{\sqrt{2}}{2}$ in the computation of 8-point FFT is w^2. Following Theorems 1 and 2, we have

$$
F_{16} = E_{16} \, \text{diag}\{F_8, F_8\} \begin{bmatrix} P_8^e \\ \cdots \\ P_o^8 \end{bmatrix}
$$

$$
= \begin{bmatrix} I_8 & D_8 \\ I_8 & -D_8 \end{bmatrix} \text{diag}\{G_0^3, G_0^3\} \, \text{diag}\{G_1^3, G_1^3\} \, \text{diag}\{G_2^3, G_2^3\} \, \text{diag}\{\widetilde{P}_8, \widetilde{P}_8\} \begin{bmatrix} P_8^e \\ \cdots \\ P_o^8 \end{bmatrix},
$$

where

$$
D_8 = \text{diag}\{1, w, w^2, w^3, -i, -\overline{w}^3, -\overline{w}^2, -\overline{w}\},
$$

and G_0^3, G_1^3, G_2^3 and $\widetilde{P}_8$ are provided above for FFT with $n = 8$.

The role of

$$
\widetilde{P}_{16} = \text{diag}\{\widetilde{P}_8, \widetilde{P}_8\} \begin{bmatrix} P_8^e \\ \cdots \\ P_o^8 \end{bmatrix}
$$

is to interchange the orders of x_j in $\mathbf{x} = [x_0, x_1, \ldots, x_{15}]^T \in \mathbb{C}^{16}$ such that

$$
\widetilde{P}_{16} \mathbf{x} = [x_0, x_8, x_4, x_{12}, x_2, x_{10}, x_6, x_{14}, x_1, x_9, x_5, x_{13}, x_3, x_{11}, x_7, x_{15}]^T \tag{4.3.12}
$$

(see Exercise 6). Denote

$$\mathbf{d} = \text{diag}\{G_2^3, G_2^3\}\widetilde{P}_{16}\mathbf{x}, \quad \mathbf{h} = \text{diag}\{G_1^3, G_1^3\}\mathbf{d},$$
$$\mathbf{g} = \text{diag}\{G_0^3, G_0^3\}\mathbf{h}, \quad \mathbf{u}(=\widehat{\mathbf{x}}) = E_{16}\mathbf{g}.$$

Then, with $w = e^{-i\frac{\pi}{8}}$, the FFT algorithm for $n = 16$ to obtain $\mathbf{u} = \widehat{\mathbf{x}} = F_{16}\mathbf{x}$ is

$$
\begin{cases}
d_0 = x_0 + x_8 \\
d_1 = x_0 - x_8 \\
d_2 = x_4 + x_{12} \\
d_3 = x_4 - x_{12} \\
d_4 = x_2 + x_{10} \\
d_5 = x_2 - x_{10} \\
d_6 = x_6 + x_{14} \\
d_7 = x_6 - x_{14} \\
d_8 = x_1 + x_9 \\
d_9 = x_1 - x_9 \\
d_{10} = x_5 + x_{13} \\
d_{11} = x_5 - x_{13} \\
d_{12} = x_3 + x_{11} \\
d_{13} = x_3 - x_{11} \\
d_{14} = x_7 + x_{15} \\
d_{15} = x_7 - x_{15}
\end{cases}
\begin{cases}
h_0 = d_0 + d_2 \\
h_1 = d_1 - id_3 \\
h_2 = d_0 - d_2 \\
h_3 = d_1 + id_3 \\
h_4 = d_4 + d_6 \\
h_5 = d_5 - id_7 \\
h_6 = d_4 - d_6 \\
h_7 = d_5 + id_7 \\
h_8 = d_8 + d_{10} \\
h_9 = d_9 - id_{11} \\
h_{10} = d_8 - d_{10} \\
h_{11} = d_9 + id_{11} \\
h_{12} = d_{12} + d_{14} \\
h_{13} = d_{13} - id_{15} \\
h_{14} = d_{12} - d_{14} \\
h_{15} = d_{13} + id_{15}
\end{cases}
\begin{cases}
g_0 = h_0 + h_4 \\
g_1 = h_1 + w^2h_5 \\
g_2 = h_2 - ih_6 \\
g_3 = h_3 - \overline{w}^2h_7 \\
g_4 = h_0 - h_4 \\
g_5 = h_1 - w^2h_5 \\
g_6 = h_2 + ih_6 \\
g_7 = h_3 + \overline{w}^2h_7 \\
g_8 = h_8 + h_{12} \\
g_9 = h_9 + w^2h_{13} \\
g_{10} = h_{10} - ih_{14} \\
g_{11} = h_{11} - \overline{w}^2h_{15} \\
g_{12} = h_8 - h_{12} \\
g_{13} = h_9 - w^2h_{13} \\
g_{14} = h_{10} + ih_{14} \\
g_{15} = h_{11} + \overline{w}^2h_{15}
\end{cases}
\begin{cases}
u_0 = g_0 + g_8 \\
u_1 = g_1 + wg_9 \\
u_2 = g_2 + w^2g_{10} \\
u_3 = g_3 + w^3g_{11} \\
u_4 = g_4 - ig_{12} \\
u_5 = g_5 - \overline{w}^3g_{13} \\
u_6 = g_6 - \overline{w}^2g_{14} \\
u_7 = g_7 - \overline{w}g_{15} \\
u_8 = g_0 - g_8 \\
u_9 = g_1 - wg_9 \\
u_{10} = g_2 - w^2g_{10} \\
u_{11} = g_3 - w^3g_{11} \\
u_{12} = g_4 + ig_{12} \\
u_{13} = g_5 + \overline{w}^3g_{13} \\
u_{14} = g_6 + \overline{w}^2g_{14} \\
u_{15} = g_7 + \overline{w}g_{15}.
\end{cases}
$$

$$\tag{4.3.13}$$

Observe that the computational complexity for computing 16-point DFT via FFT consists of 17 complex multiplications and 64 additions. The algorithm in (4.3.13) is ready to be implemented easily. ∎

The signal flow charts for 4-point FFT and 8-point FFT are provided in Figs. 4.4 and 4.5, respectively.
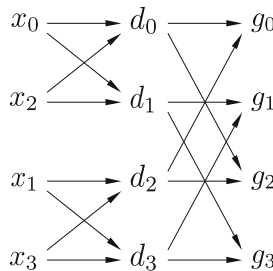
Fig. 4.4 FFT Signal flow chart for $n = 4$
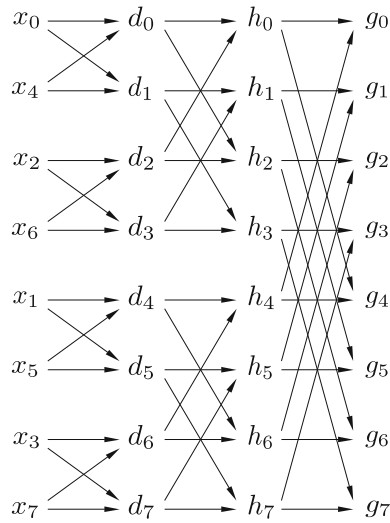
Fig. 4.5 FFT signal flow chart for $n = 8$

Exercises

Exercise 1 For $n = 2^m$, where m is an integer, verify that the number of operations required to compute the n-point DFT of any vector $\mathbf{x} \in \mathbb{C}^n$ is of order $O(mn) = O(n \log n)$, when the FFT procedure, as described in (4.3.6) of Theorem 2, is followed.

Exercise 2 Derive the formula (4.3.7).

Exercise 3 Verify the computational scheme (4.3.9).

Exercise 4 Verify the validity of the 8×8 permutation matrix (4.3.10).

Exercise 5 Verify the computational scheme (4.3.11).

Exercise 6 Write out the permutation matrix $\widetilde{P}_{16}$ and verify (4.3.12).

Exercise 7 Fill in the details in the derivation of the computational scheme of the 16-point FFT in (4.3.13).

Exercise 8 Attempt to draw the signal flow chart for the 16-point FFT by applying (4.3.13).

4.4 Fast Discrete Cosine Transform

The DCT introduced in Sect. 4.2 (see Definition 1 on p.183) is called DCT-II. It is the DCT used in Matlab and certain compression standards, such as JPEG image compression and MPEG video compression to be discussed in the next chapter

(see Sect. 5.4). Recall that this DCT is derived from the DFT in (4.2.1)–(4.2.10) in Sect. 4.2, and observe that the derivation of this DCT $C_n^{II} = C_n$ in (4.2.10) on p.183 from the DFT F_{2n} can be formulated as the product of the matrix F_{2n}, a matrix P of the $4n$th roots of unity, and a permutation matrix $Q = Q_{2n \times n}$, as follows:

$$C_n^{II} = C_n = P^* F_{2n} Q, \tag{4.4.1}$$

where F_{2n} is given by (4.1.8) on p.174 with n replaced by $2n$,

$$P = P_{2n \times n} = \frac{1}{2\sqrt{2n}} \begin{bmatrix} \sqrt{2} & & & \\ & \omega_n & & \\ & & \ddots & \\ & & & \omega_n^{n-1} \\ 0 & \cdots\cdots & & 0 \\ & & & \overline{\omega}_n^{n-1} \\ & & \cdot^{\cdot^{\cdot}} & \\ 0 & \overline{\omega}_n & & \end{bmatrix}_{2n \times n} \tag{4.4.2}$$

with

$$\omega_n = e^{i\frac{\pi}{2n}}, \tag{4.4.3}$$

and

$$Q = Q_{2n \times n} = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \\ & & & 1 \\ & & \cdot^{\cdot^{\cdot}} & \\ & 1 & & \\ 1 & & & \end{bmatrix}_{2n \times n}, \tag{4.4.4}$$

where the zero entries in P and Q are not displayed.

There are other types of DCT that are commonly used, including DCT-I, DCT-III, and DCT-IV. For convenience, we introduce the notation

$$b(k) = \begin{cases} \frac{1}{\sqrt{2}}, & \text{for } = 0, n, \\ 1, & \text{for } \le k \le n-1, \\ 0, & \text{for } < 0 \text{ or } k > n, \end{cases} \tag{4.4.5}$$

for formulating the DCT matrices of DCT-I, DCT-II, and DCT-III, as follows.

(i) DCT-I is defined by the $(n+1) \times (n+1)$ orthogonal matrix

$$C_{n+1}^{I} = \left[c_{n+1}^{I}(k, \ell) \right]$$

with the (k, ℓ)-entry

$$c_{n+1}^{I}(k, \ell) = b(k)b(\ell)\sqrt{\frac{2}{n}} \cos \frac{k\ell\pi}{n}, \tag{4.4.6}$$

where $k, \ell = 0, \ldots, n$.

(ii) DCT-II is defined by the $n \times n$ orthogonal matrix

$$C_n^{II} = \left[c_n^{II}(k, \ell) \right]$$

with the (k, ℓ)-entry

$$c_n^{II}(k, \ell) = b(k)\sqrt{\frac{2}{n}} \cos \frac{k(\ell + \frac{1}{2})\pi}{n}, \tag{4.4.7}$$

where $k, \ell = 0, \ldots, n - 1$ (see (4.2.9) on p.174).

(iii) DCT-III is the transpose of C_n^{II}, defined by

$$C_n^{III} = \left[c_n^{III}(k, \ell) \right]$$

with the (k, ℓ)-entry

$$c_n^{III}(k, \ell) = b(\ell)\sqrt{\frac{2}{n}} \cos \frac{(k + \frac{1}{2})\ell\pi}{n}, \tag{4.4.8}$$

where $k, \ell = 0, \ldots, n - 1$ (note that $c_n^{III}(k, \ell) = c_n^{II}(\ell, k)$).

(iv) DCT-IV is defined by the $n \times n$ orthogonal matrix

$$C_n^{IV} = \left[c_n^{IV}(k, \ell) \right]$$

with the (k, ℓ)-entry

$$c_n^{IV}(k, \ell) = \sqrt{\frac{2}{n}} \cos \frac{(k + \frac{1}{2})(\ell + \frac{1}{2})\pi}{n}, \tag{4.4.9}$$

where $k, \ell = 0, \ldots, n - 1$.

Remark 1 Since each of the matrices C_{n+1}^{I}, C_n^{II}, C_n^{III}, and C_n^{IV} is an orthogonal matrix (see Exercises 1–4), the corresponding inverse matrix is its transpose. Observe that since C_{n+1}^{I} and C_n^{IV} are symmetric matrices, they are their own inverses, namely:

$$\left(C_{n+1}^{I}\right)^2 = I_{n+1}, \quad \left(C_n^{IV}\right)^2 = I_n,$$

where I_{n+1} and I_n are identity matrices of dimensions $n+1$ and n, respectively. Also, as mentioned above, since C_n^{III} is the transpose of C_n^{II}, we have $\left(C_n^{II}\right)$ $\left(C_n^{III}\right) = I_n$. ■

In (4.4.1), we give the matrix representation of C_n^{II} in terms of the DFT matrix F_{2n}. Since C_n^{III} is the transpose of C_n^{II}, it can be written as

$$C_n^{III} = Q^T F_{2n} \overline{P}, \tag{4.4.10}$$

in view of the fact that the matrix F_{2n} is symmetric (see (4.1.8) on p.183). As to C_{n+1}^{I} and C_n^{IV}, we have

$$C_{n+1}^{I} = \frac{1}{2\sqrt{2n}} U^T F_{2n} U, \tag{4.4.11}$$

where

$$U = U_{2n \times (n+1)} = \begin{bmatrix} \sqrt{2} & & & & & \\ & 1 & & & & \\ & & \ddots & & & \\ & & & 1 & & \\ & & & 0 & \sqrt{2} & \\ & & & 1 & & \\ & & & & \ddots & \\ 0 & & & & 1 \end{bmatrix}; \tag{4.4.12}$$

and

$$C_n^{IV} = \frac{1}{2\sqrt{2n}} e^{-i\pi/4n} R^T F_{2n} R, \tag{4.4.13}$$

where

$$R = R_{2n \times n} = \begin{bmatrix} 1 & & & & \\ & \overline{\omega}_n & & & \\ & & \ddots & & \\ & & & \overline{\omega}_n^{n-2} & \\ & & & & \overline{\omega}_n^{n-1} \\ 0 & \cdots\cdots & 0 & & i \\ & & & \omega_n^{n-1} & 0 \\ & & & \ddots & \\ & & \ddots & & \\ \omega_n & & & & \end{bmatrix} \tag{4.4.14}$$

with $\omega_n = e^{i\frac{\pi}{2n}}$.

The importance of the matrix-product formulation of the DCT, in (4.4.1), (4.4.10), (4.4.11), and (4.4.13), is that the FFT, as introduced in Sect. 4.3, can be applied to achieve fast computation of the DCT (called FDCT), with $O(n \log n)$ operations, as opposed to $O(n^2)$ operations, provided that $n = 2^m$ for some integer $m > 0$. The reason is that multiplication by the matrices P, Q, U and R (in (4.4.2), (4.4.4), (4.4.12) and (4.4.14), respectively), or by their inverses, does not increase the order $O(n \log n)$ of operations (see Exercise 8).

DCT-II is more often used than DCT-I and DCT-IV. For example, for JPEG image and MPEG video compressions, the digital image is divided into 8×8 pixel blocks and 8-point DCT-II is applied to both the horizontal and vertical directions of each block. In Chap. 5, we will describe the general compression scheme in Sect. 5.3 and JPEG image compression standard in Sect. 5.4. In Figs. 4.6, 4.7, 4.8 we include DCT-II via FDCT computations without going into any details. The interested reader is referred to the vast literature on JPEG image compression, the topic to be studied in the next chapter (see Exercises 9-11).

Exercises

Exercise 1 Verify that matrix C_{n+1}^I with entries given by (4.4.6) is orthogonal for $n = 3$.

Exercise 2 Verify that matrix C_n^{IV} with entries given by (4.4.9) is orthogonal for $n = 3$ and $n = 4$.
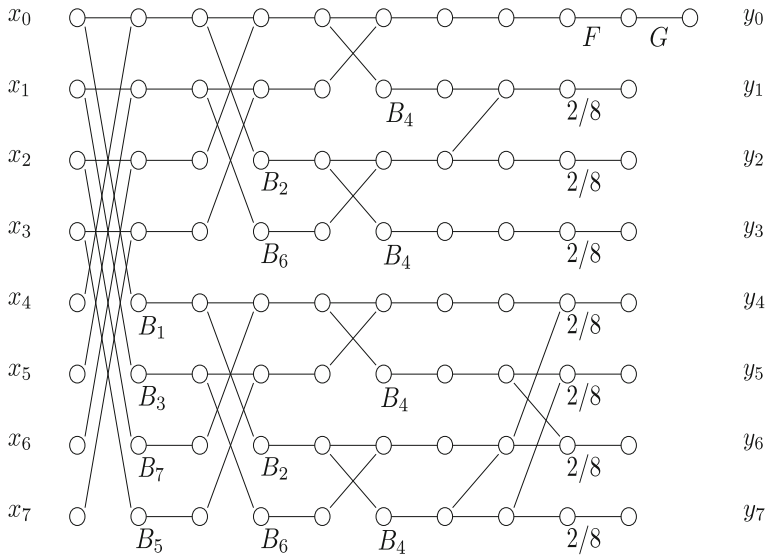
Fig. 4.6 Lee's fast DCT implementation
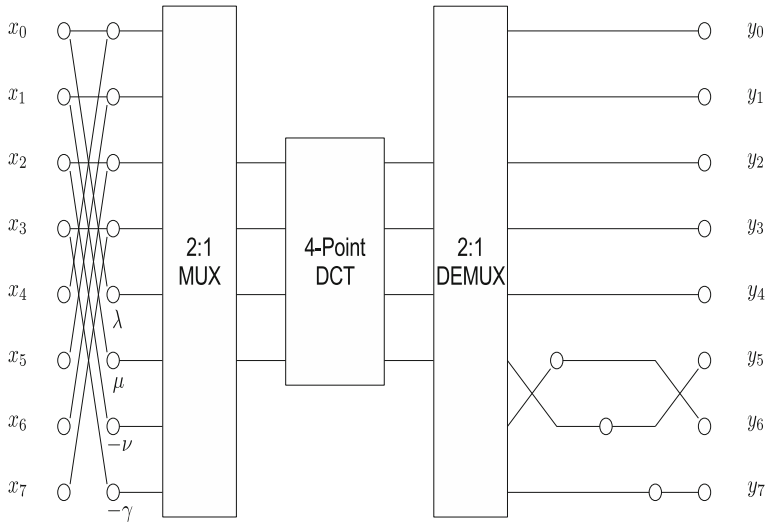
Fig. 4.7 Hou's Fast DCT implementation
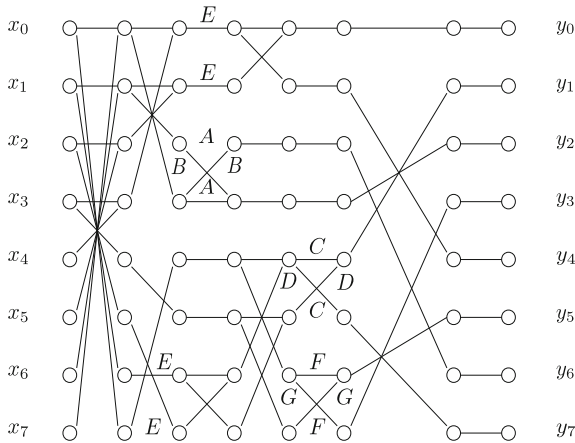
Fig. 4.8 Fast DCT implementation of Wang, Suehiro and Hatori

Exercise 3 Prove that matrix C_{n+1}^{I} is an orthogonal matrix for all positive integers n, by showing that $\left(C_{n+1}^{I}\right)^{2} = I_{n+1}$.

Exercise 4 Prove that matrix C_{n}^{IV} is an orthogonal matrix for all positive integers n, by showing that $\left(C_{n}^{IV}\right)^{2} = I_{n}$.

Exercise 5 Verify for $n = 3$ and $n = 4$, that matrix C_{n}^{II} with entries given by (4.4.7) satisfies (4.4.1), where $P = P_{2n \times n}$ and $Q = Q_{2n \times n}$ are defined by (4.4.2) and (4.4.4), respectively.

Exercise 6 Verify for $n = 3$, that matrix C_{n+1}^I with entries given by (4.4.6) satisfies (4.4.11), where $U = U_{2n \times (n+1)}$ is defined by (4.4.12).

Exercise 7 Verify for $n = 3$ and $n = 4$, that C_n^{IV} with entries given by (4.4.9) satisfies (4.4.13), where $R = R_{2n \times n}$ is defined by (4.4.14).

Exercise 8 Apply the complexity count of FFT from Sect. 4.3 to show that the FDCT has $O(n \log n)$ operations for C_{n+1}^I, C_n^{II}, and C_n^{IV}. (Note that since C_n^{III} is the transpose of C_n^{II}, the FDCT for C_n^{III} has the same number of operations as C_n^{II}.)

Exercise 9 Verify the correctness of Lee's FDCT implementation (in Fig. 4.6) for $n = 8$.

Exercise 10 Verify the correctness of Hou's FDCT implementation (in Fig. 4.7) for $n = 8$.

Exercise 11 Verify the correctness of FDCT implementation (in Fig. 4.8) by Wang, Suehiro and Hatori, for $n = 8$.

Chapter 5
Data Compression
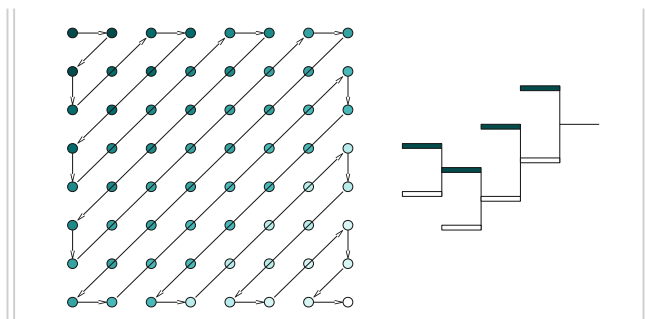
The difference between data reduction, including data dimensionality reduction studied in Chap. 3, and the topic of data compression to be investigated in this chapter is that compressed data must be recoverable, at least approximately. The most commonly used representation of data (particularly compressed data) for communication and storage is a string of numbers consisting only of zeros, 0's, and ones, 1's, without using any punctuation. This string of 0's and 1's is called a "binary code" of the data. For the recovery of the data, whether compressed or not, the binary code must be decipherable by referring to the corresponding code-table. The length of the binary code depends on coding efficiency, which is governed by the "entropy" of the source data. On the other hand, by manipulating the source data in a clever way, the entropy could be decreased, and hence allows for shorter binary codes. In addition, if recovery of a fairly good approximation of the source data is sufficient, the entropy could be significantly decreased to allow for very high compression ratios. This is called "lossy" (or irreversible) compression. Lossy compression of photographs (particularly, digital images) and videos are most effective, since the human eye is not very sensitive to the high-frequency contents of imagery data.

To quantify the efficiency of coding schemes, the notion of "entropy" is introduced in Sect. 5.1, in terms of the probability distribution of the data. In addition to establishing the theoretical results, examples are given to illustrate the feasibility of decreasing this governing quantity. Binary codes are studied in the second section. While Kraft's inequality governs the length of individual code-words, the average code length is shown to be bounded below by the entropy. The most commonly used (entropy) coding algorithm, called Huffman coding, is discussed in some details in Sect. 5.2. It is also demonstrated in this section that the Huffman code does not exceed the upper bound of the Noiseless Coding Theorem. The topic of data compression schemes is discussed in Sect. 5.3, and specifically, the application of DCT, studied in Chap. 4, is applied to segments (also called blocks or tiles) of the data sequence to facilitate efficient computation and memory saving, particularly for long

data sequences. Unfortunately, if the DCT segments are quantized, the inverse DCT, again applied to individual segments, reveals "blocky artifacts", called " quantization noise". To suppress the quantization noise, the data segments can be pre-processed in such a way that the DCT applied to each segment is changed to a "windowed DCT" of the same segment, along with some data in its two adjacent segments. Since DCT is an orthogonal matrix and the windowed DCT extends the original DCT to operate on some data from the adjacent segments, the windowed DCT is also called a lapped orthogonal transform, with abbreviation "LOT". The more general "lapped transform", without the orthogonality restriction, is also introduced in Sect. 5.3 for the purpose of constructing the general LOT, while the notion of "dual lapped transform" is also introduced to yield the inverse of the corresponding LOT, via data post-processing. Digital image compression, and in particular the JPEG compression standard, is the central theme of investigation in Sect. 5.1, that includes the study of color formats and extension to video compression.

5.1 Entropy

Information Theory is an area of Applied Mathematics which is concerned with such topics as the quantification, coding, communication, and storage of information. Here, the term "information" should not be confused with the word "meaning", since in Information Theory, "pure nonsense" could be considered as an information source. In other words, "information" should not be measured in terms of "what is to be said", but only of "what could be said". As Claude Shannon, the father of Information Theory, emphasized in his 1948 pioneering paper, the semantic aspects of the communication of information are irrelevant to mathematical and engineering considerations.

On the other hand, coding of an information source is essential for communication and storage of the information. Here, the term "coding" consists of two operations: "encoding" of the information source to facilitate effective communication and storage, and "decoding" to recover the original information source. In other words, decoding is the inverse operation of encoding. In this chapter, we only consider "binary coding", meaning that the encoded information, called a binary code, is a (finite) "sequence" of numbers consisting only of zeros, 0's, and ones, 1's; and that all "code-words" and "instructions" that constitute this so-called sequence can be identified from some "code-table", even though no punctuation are used in this "sequence" to separate them. Furthermore, all elements of the set of the information source are represented as code-words in the code-table. Here, the term "sequence" is in quotation, since no "commas" are allowed, as opposed to the usual mathematical definition of sequences.

Hence, if each element of the given set of information source is assigned some non-negative integer, all of such integers with the "code-words" contained in some "code-table", then decoding to receive or recover the original information source from the code is possible by using the code-table. To quantify the integer representation

of (the elements of the set of) the information source, the unit "bit", coined by John Tukey as the abbreviation of "binary digits" is used. For example, if the largest (non-negative) integer used in the integer representation of the information source is 255, we say that the source is an 8-bit information (since the binary representation of 255 is 11111111, a string of 8 ones). In other words, one bit (or 1b) is one 0 or 1. In addition, the measurement in the unit of "bytes" is also used. One byte (1B) is equal to 8 bits (1B = 8b).

When a binary code is stored in some memory device (such as hard disk, flash memory, or server), the length of the "sequence" of 0's and 1's is called the "file size" of the (compressed) data. When the binary code is transmitted (such as in broadcasting or video streaming), the length of the sequence of 0's and 1's is called the length of the bit-stream, and the speed of transmission is called the "bit-rate". While file sizes are measured in kilo-bytes, mega-bytes, giga-bytes, and tera-bytes (or kB, MB, GB, and TB, respectively), bit-rates are measured in kilo-bits per second, mega-bits per second, and giga-bits per second (or kb/s, Mb/s, Gb/s, respectively).

A typical example is an 8-bit gray-scale digital image, with the "intensity" of each pixel (considered as an element of the set of the information source, which is the image) being calibrated in increasing integer steps from 0 to 255, with 0 representing "black" (or no light) and 255 representing "white", while for $j = 1, \ldots, 254$, the increase in intensity yields increasingly lighter shades (of gray). Another example is a 24-bit color digital image. But instead of using 24 bits in the binary expression of the largest integer used in the integer representation of the information source (which is a color image), the convention is to assign 8-bits to each of the three primary color components, "red" (R), "green" (G), and "blue" (B), in that a triple (i, j, k) of integers, with each of i, j, and k, ranging from 0 to 255, for (R, G, B) represents increasing intensities of the red, green, and blue (visible) lights, respectively. Recall that as primary additive colors, addition of different intensities for R, G, B (that is, different values of i, j, k) yield $2^8 \times 2^8 \times 2^8 = 2^{24}$ colors of various shades. In particular, a gray-scale digital image is obtained by setting $i = j = k$, where $0 \leq i \leq 255$. Therefore, the term "24-bit color" is justified.

It must be understood that the meaning of a 12-bit novel only indicates an upper bound of the number of words in the novel. The actual encoded file size is significantly larger, and often quantified in megabytes. The typical compressed file size of a novel is about 1 MB. Similarly, by a 24-bit color picture, we only mean that the quality (in terms of shades of color) is limited to 24-bit. The file size of a JPEG compressed image usually exceeds several kilo-bytes and occasionally even over 1 MB, depending on the image resolution (that is, the number of pixels).

In the above examples, it will be clear that the notion of "probability", in the sense of percentages of equal pixel values for a digital image and percentages of the same word being used in the novel, should play an important role in information coding. For instance, in a picture with blue sky and blue water in the background, the percentages of RGB pixels with values of $(0, 0, k)$, where $50 \leq k \leq 150$, are much higher than those with $(i, j, 0)$ for all $0 \leq i, j \leq 255$. Also, for the novel example mentioned above, the frequency of occurrence of the word "the" is much higher than just about all of the other words in the novel. In fact, according to some study, less

than half of the vocabulary used in a typical novel constitutes over 80 % of all the words in the book.

Furthermore, it should be emphasized that the "entropy" of the information source (to be defined next in terms of the probabilities of occurrence as discussed above) often decreases when some suitable mathematical transformation is applied to the integer representation of the information source. Typical transforms include RLE and DPCM, to be discussed briefly in Example 1 on p.210 and again in Sects. 5.3 and 5.4 on digital image compression. Since the encoded file size and length of an encoded bit-stream are governed by the entropy, it is important to understand this concept well.

Definition 1 $\boxed{\textbf{Entropy}}$ *Let $X_n = \{x_1, \ldots, x_n\}$ be an information source (or some mathematical transformation of a given information source). Let $Z = \{z_1, \ldots, z_m\}$ be the subset of all distinct elements of the set X_n. Corresponding to each $z_j \in Z$, $j = 1, \ldots, m$, let p_j denote its probability of occurrence in X_n. Then $\mathcal{P} = \{p_1, \ldots, p_m\}$ is a discrete probability distribution, and the function*

$$H(\mathcal{P}) = H(p_1, \ldots, p_m) = \sum_{j=1}^{m} p_j \log_2 \frac{1}{p_j} \tag{5.1.1}$$

is called the entropy of the probability distribution $\mathcal{P}$, or more specifically, the entropy of the pair $(X_n, \mathcal{P})$.

Remark 1 For most applications, since the $\#X_n = n$ is very large and since many (or even most) information sources of the same application are quite similar, the same discrete probability distribution $\mathcal{P}$ is used for all such information sources. This $\mathcal{P}$ is usually obtained from large volumes of previous experiments. Of course, the precise values of $p_1, \ldots, p_m$ that constitute $\mathcal{P}$ can be defined (and computed) in terms of the "histogram" of X_n; namely,

$$p_j = \frac{1}{n} \#\{x_i \in X_n : x_i = z_j\}, \tag{5.1.2}$$

for each $j = 1, \ldots, m$.   ∎

Example 1 Let the information source be the data set $X = \{0, 1, \ldots, 255\}$. Introduce some transformation of X to another set Y that facilitates binary coding which must be reversible.

Solution The set X can be written as $X = \{x_0, \ldots, x_{255}\}$ with $x_j = j$ for $j = 0, \ldots, 255$. Since the binary representation of 255 is a string of 8 ones; namely $x_{255} = 11111111$, and since no commas are allowed to separate x_j from its neighbor x_{j+1}, the straight forward and naive way is to assign 8 bits to each of $j = 0, \ldots, 255$; namely, $x_0 = 00000000$, $x_1 = 00000001, \ldots, x_{255} = 11111111$. Hence, the code is $00000000000000010 \cdots 011111111$, which has length of $256 \times 8 = 2048$ bits.

On the other hand, if we set $y_0 = 0$ and

$$y_j = x_j - x_{j-1} - 1, \text{ for } j = 1, \ldots, 255, \qquad (5.1.3)$$

then we have $y_j = 0$, for all $j = 0, \ldots, 255$. Therefore, the code for $Y = \{y_0, \ldots, y_{255}\} = \{0, \ldots, 0\}$ is simply $0 \cdots 0$, a string of 256 zeros. The transformation from X to Y is reversible, since, for $y_j = 0$, $j = 0, \ldots, 255$, we have

$$x_j = y_j + x_{j-1} + 1, \ j = 1, \ldots, 255, \qquad (5.1.4)$$

with initial condition $x_0 = 0$. Of course the formula (5.1.4) along with $x_0 = 0$ must be coded, but this requires only a few bits. In addition, instead of coding a string of 256 zeros in (5.1.3) with $y_0 = 0$, one may "encode" the size of a block of 256 zeros. This is called "run-length encoding (RLE)". Also, the code of the transformation in (5.1.3) is called the "differential pulse code modulation (DPCM)". Both RLE and DPCM are commonly used, and are part of the JPEG coding scheme for image compression. We will elaborate on RLE and DPCM in Sect. 5.4 later. ∎

Example 2 As a continuation of Example 1 on p. 210 compute the entropy of the information source $X_n = \{x_0, \ldots, x_{255}\}$ with $x_j = j$ and $n = 256$ and the DPCM transformation $Y_{256} = \{y_0, \ldots, y_{255}\}$, with $y_0 = 0$ and

$$y_j = x_j - x_{j-1} - 1, \ j = 1, \ldots, 255.$$

Solution For X_{256}, since all its elements are distinct, we have $p_0 = \cdots = p_{255} = \frac{1}{256}$, and hence, the entropy of $(X_{256}, \mathcal{P})$ is given by

$$H(p_0, \ldots, p_{255}) = \sum_{j=0}^{255} p_j \log_2 \frac{1}{p_j} = \frac{1}{256} \sum_{j=0}^{255} \log_2 256 = 8.$$

On the other hand, since $Y_{256} = \{0, \ldots, 0\}$, the probability distribution is $\mathcal{P} = \{1\}$, so that

$$H(1) = 0. \qquad ∎$$

In the following, we show that with the given probability distribution as weights to define the weighted average of the logarithm of another probability distribution $\mathcal{Q}$, then $\mathcal{P}$ is the optimal choice among all $\mathcal{Q}$.

Theorem 1 *Let $\mathcal{P} = \{p_1, \ldots, p_m\}$ be a given discrete probability distribution with entropy $H(\mathcal{P})$ defined by (5.1.1) of Definition 1. Then for all probability distributions $\mathcal{Q} = \{q_1, \ldots, q_m\}$; that is, $0 \le q_j \le 1$ for $j = 1, \ldots, m$, and $q_1 + \cdots + q_m = 1$, the quantity*

$$G(q_1, \ldots, q_m) = \sum_{j=1}^{m} p_j \log_2 \frac{1}{q_j}$$

satisfies

$$G(q_1, \ldots, q_m) \geq G(p_1, \ldots, p_m) = H(\mathcal{P}). \tag{5.1.5}$$

Furthermore, $G(q_1, \ldots, q_m) = H(\mathcal{P})$ *in* (5.1.5), *if and only if* $q_1 = p_1, \ldots, q_m = p_m$.

The result in (5.1.5) can be reformulated as

$$H(\mathcal{P}) = \inf_{\mathcal{Q}} G(\mathcal{Q}), \tag{5.1.6}$$

where inf (which stands for "infimum") is interpreted as the greatest lower bound over all probability distributions $\mathcal{Q}$.

| **Proof of Theorem 1** | To prove (5.1.5), let us consider the constraint optimization problem

$$\min\{G(x_1, \ldots, x_m) : \sum_{j=1}^{m} x_j = 1\}. \tag{5.1.7}$$

By introducing a parameter λ, called the Lagrange multiplier, the problem (5.1.7) can be solved by applying the "method of Lagrange multipliers", namely:

$$\min_{x_1, \ldots, x_m, \lambda} G_\lambda(x_1, \ldots, x_m, \lambda),$$

where

$$G_\lambda(x_1, \ldots, x_m, \lambda) = G(x_1, \ldots, x_m) + \lambda\Big(\sum_{j=1}^{m} x_j - 1\Big). \tag{5.1.8}$$

To find the local minima of (5.1.8), we may compute the partial derivatives of G_λ with respect to each of $x_1, \ldots, x_m, \lambda$ and set them to be zero, yielding:

$$0 = \frac{\partial}{\partial x_\ell} G_\lambda = \frac{\partial}{\partial x_\ell} p_\ell \log_2 \frac{1}{x_\ell} + \lambda$$
$$= p_\ell \left(- \frac{1}{x_\ell \ln 2} \right) + \lambda, \quad \ell = 1, \ldots, m, \tag{5.1.9}$$

and

$$0 = \frac{\partial}{\partial \lambda} G_\lambda = \sum_{j=1}^{m} x_j - 1. \tag{5.1.10}$$

By (5.1.9), we have

$$\frac{p_\ell}{x_\ell} = \lambda \ln 2, \quad \ell = 1, \ldots, m, \tag{5.1.11}$$

so that, in view of (5.1.10),

$$1 = \sum_{\ell=1}^{m} p_\ell = \lambda \ln 2 \sum_{\ell=1}^{m} x_\ell = \lambda \ln 2$$

or $\lambda = \frac{1}{\ln 2}$; and it follows from (5.1.11) that

$$\frac{p_\ell}{x_\ell} = \frac{\ln 2}{\ln 2} = 1,$$

or equivalently $x_\ell = p_\ell$ (i.e. $q_\ell = p_\ell$), for $\ell = 1, \ldots, m$. This completes the proof of Theorem 1 or (5.1.6). ∎

In the following, we list two useful properties of the entropy function.

Theorem 2 *The entropy function $H(\mathcal{P})$ defined on the collection of all discrete probability distributions $\mathcal{P}$ satisfies*

$$0 \le H(\mathcal{P}) \le \log_2 m, \tag{5.1.12}$$

where m denotes the number of positive masses of $\mathcal{P}$; namely, $\mathcal{P} = \{p_1, \ldots, p_m\}$ with $p_j > 0$ for all $j = 1, \ldots, m$. Furthermore, $H(x_1, \ldots, x_m)$ is a continuous function on the simplex

$$D_m = \{(x_1, \ldots, x_m) : \sum_{j=1}^{m} x_j = 1, \ 0 \le x_1, \ldots, x_m \le 1\},$$

and has continuous partial derivatives in the interior of D.

We remark that the lower bound in (5.1.12) is attained at the discrete probability distribution $\mathcal{P} = \{1\}$ with a single positive mass, while the upper bound in (5.1.12) is attained at the discrete probability distribution $\mathcal{P} = \{\frac{1}{m}, \ldots, \frac{1}{m}\}$ with m equal masses $p_1 = \cdots = p_m = \frac{1}{m}$ (see Example 2 on p.211).

Proof of Theorem 2 That the lower bound in (5.1.12) is valid follows from the fact that $\log_2(\frac{1}{p_j}) \ge 0$ for $0 < p_j \le 1$, since

$$H(\mathcal{P}) = \sum_{j=1}^{m} p_j \log_2 \frac{1}{p_j} \ge 0.$$

To derive the upper bound in (5.1.12), we may apply the method of Lagrange multipliers introduced in the proof of Theorem 1 , by considering the function

$$G(x_1, \ldots, x_m, \lambda) = \sum_{j=1}^{m} x_j \log_2 \frac{1}{x_j} + \lambda \left(\sum_{j=1}^{m} x_j - 1 \right),$$

and setting all first order partial derivatives of $G(x_1, \ldots, x_m, \lambda)$ to be zero; namely,

$$
\begin{aligned}
0 &= \frac{\partial}{\partial x_\ell} G = \frac{\partial}{\partial x_\ell} \left(x_\ell \log_2 \frac{1}{x_\ell} \right) + \lambda \\
&= \log_2 \frac{1}{x_\ell} + x_\ell \left(-\frac{1}{x_\ell \ln 2} \right) + \lambda, \quad \ell = 1, \ldots, m,
\end{aligned}
\tag{5.1.13}
$$

and

$$
0 = \frac{\partial}{\partial \lambda} G = \sum_{j=1}^{m} x_j - 1.
\tag{5.1.14}
$$

From (5.1.13), we have, for all $\ell = 1, \ldots, m$,

$$
\log_2 x_\ell = \lambda - \frac{1}{\ln 2}
$$

or $x_1 = x_2 = \cdots = x_m = 2^{(\lambda - \frac{1}{\ln 2})}$, so that (5.1.14) implies that

$$
x_1 = \cdots = x_m = \frac{1}{m}.
$$

Hence,

$$
\begin{aligned}
H(x_1, \ldots, x_m) &= \sum_{j=1}^{m} x_j \log_2 \frac{1}{x_j} \\
&= m \frac{1}{m} \log_2 \frac{1}{1/m} = \log_2 m,
\end{aligned}
$$

which is the upper bound in (5.1.12).

To complete the proof of Theorem 2, we observe that the function $f(x) = x \log_2 x$ is continuous on the closed interval $[0, 1]$ and has continuous derivatives for $0 < x \leq 1$. ∎

For discrete probability distributions $\mathcal{P} = \{p_1, \ldots, p_m\}$ with a large number m of positive masses, the following result allows the partitioning of $\mathcal{P}$ into arbitrarily many disjoint subsets of probability distributions (obtained after appropriate normalization) to facilitate computing the entropy $H(\mathcal{P})$ of $\mathcal{P}$. To facilitate the statement of this result, let $n < m$ be arbitrarily chosen and write

$$
\{p_1, \ldots, p_m\} = \{p_{k_0+1}, \ldots, p_{k_1}; \ p_{k_1+1}, \ldots, p_{k_2}; \ldots; \ p_{k_{n-1}+1}, \ldots, p_{k_n}\},
\tag{5.1.15}
$$

with $0 = k_0 < k_1 < \cdots < k_n = m$, where $k_1, \ldots, k_{n-1}$ are arbitrarily selected so that

$$\sum_{j=1}^{k_{\ell+1}-k_\ell} p_{k_\ell+j} = a_{\ell+1} > 0 \qquad (5.1.16)$$

for each $\ell = 0, \ldots, n-1$. Observe that

$$\mathcal{A} = \{a_1, \ldots, a_n\}, \qquad (5.1.17)$$

and all of

$$\widetilde{\mathcal{P}}_\ell = \{p_{k_\ell+1}/a_{\ell+1},\, p_{k_\ell+2}/a_{\ell+1},\, \ldots,\, p_{k_{\ell+1}-1}/a_{\ell+1},\, p_{k_{\ell+1}}/a_{\ell+1}\} \qquad (5.1.18)$$

for $\ell = 0, \ldots, n$, are discrete probability distributions (see Exercise 7).

Theorem 3 *Let $\mathcal{P} = \{p_1, \ldots, p_m\}$ be a discrete probability distribution, partitioned arbitrarily as in (5.1.15) such that (5.1.16) itholds. Also, let $\mathcal{A}, \widetilde{\mathcal{P}}_0, \ldots, \widetilde{\mathcal{P}}_{n-1}$ be the discrete probability distributions defined by (5.1.17) and (5.1.18). Then the entropy of $\mathcal{P}$ can be partitioned as follows:*

$$H(\mathcal{P}) = H(\mathcal{A}) + \sum_{\ell=0}^{n-1} a_{\ell+1} H(\widetilde{\mathcal{P}}_\ell). \qquad (5.1.19)$$

Proof To derive the formula (5.1.19), we observe, for each $\ell = 0, \ldots, n-1$, that

$$a_{\ell+1} H(\widetilde{\mathcal{P}}_\ell) = a_{\ell+1} \sum_{j=1}^{k_{\ell+1}-k_\ell} \frac{p_{k_\ell+j}}{a_{\ell+1}} \log_2 \frac{a_{\ell+1}}{p_{k_\ell+1}}$$

$$= \sum_{j=1}^{k_{\ell+1}-k_\ell} p_{k_\ell+j} \log_2 \frac{1}{p_{k_\ell+j}} - \left(\sum_{j=1}^{k_{\ell+1}-k_\ell} p_{k_\ell+j} \right) \log_2 \frac{1}{a_{\ell+1}}$$

$$= \sum_{j=1}^{k_{\ell+1}-k_\ell} p_{k_\ell+j} \log_2 \frac{1}{p_{k_\ell+j}} - a_{\ell+1} \log_2 \frac{1}{a_{\ell+1}}.$$

Therefore, summing both sides of the above formula over all $\ell = 0, \ldots, n-1$ yields

$$\sum_{\ell=0}^{n-1} a_{\ell+1} H(\widetilde{\mathcal{P}}_\ell) = \sum_{\ell=0}^{n-1} \sum_{j=1}^{k_{\ell+1}-k_\ell} p_{k_\ell+j} \log_2 \frac{1}{p_{k_\ell+j}} - H(\mathcal{A}) = H(\mathcal{P}) - H(\mathcal{A}),$$

completing the proof of (5.1.19). ∎

Example 3 Compute the entropy of the discrete probability distribution

$$\mathcal{P} = \left\{\frac{1}{6}, \frac{1}{3}, \frac{1}{6}, \frac{1}{12}, \frac{1}{12}, \frac{1}{6}\right\}$$

by applying Theorem 3.

Solution Write $\mathcal{P} = \left\{\frac{1}{12}, \frac{1}{12}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{3}\right\}$. Then with $m = 6$ and $n = 3$ in (5.1.15), and

$$a_1 = \frac{1}{12} + \frac{1}{12} = \frac{1}{6};$$

$$a_2 = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6};$$

$$a_3 = \frac{1}{3} = \frac{2}{6},$$

we have

$$H(\mathcal{P}) = H(a_1, a_2, a_3) + a_1 H\left(\frac{1}{12a_1}, \frac{1}{12a_1}\right) + a_2 H\left(\frac{1}{6a_2}, \frac{1}{6a_2}, \frac{1}{6a_2}\right)$$

$$\quad + a_3 H\left(\frac{1}{3a_3}\right)$$

$$= H\left(\frac{1}{6}, \frac{3}{6}, \frac{2}{6}\right) + \frac{1}{6} H\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{3}{6} H\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right) + \frac{2}{6} H(1)$$

$$= \left(\frac{1}{6} \log_2 6 + \frac{3}{6} \log_2 \frac{6}{3} + \frac{2}{6} \log_2 \frac{6}{2}\right) + \frac{1}{6} 2\left(\frac{1}{2} \log_2 2\right)$$

$$\quad + \frac{3}{6} 3\left(\frac{1}{3} \log_2 3\right) + 0$$

$$= \left(\frac{1}{6} \log_2 3 + \frac{1}{6} + \frac{3}{6} + \frac{2}{6} \log_2 3\right) + \frac{1}{6} + \frac{3}{6} \log_2 3$$

$$= \frac{5}{6} + \log_2 3.$$

■

<div align="center">

Exercises

</div>

Exercise 1 Let $\mathcal{P}_5 = \{p_1, \ldots, p_5\}$ be a discrete probability distribution, where p_j denotes the probability of occurrence of the integer j in a given information source X_n for $j = 1, \ldots, 5$. Compute $\mathcal{P}_5 = \mathcal{P}_5(X_n)$ for each of the following information sources in terms of the histogram of X_n.

(a) $X_8 = \{2, 2, 3, 5, 4, 1, 1, 1\}$.
(b) $X_{14} = \{5, 5, 3, 4, 2, 5, 4, 4, 3, 2, 1, 3, 5, 1\}$.
(c) $X_{16} = \{1, 3, 2, 1, 3, 2, 1, 3, 3, 3, 4, 5, 4, 3, 3, 1\}$.

Exercise 2 As a continuation of Exercise 1, compute the entropy of each of the information sources X_8, X_{14}, X_{16} (or more precisely, the entropy $H(\mathcal{P}_5)$ for each information source).

Exercise 3 Let $\mathcal{P}_8 = \{p_1, \ldots, p_8\}$ be a discrete probability distribution, where p_j denotes the probability of occurrence of the source word a_j in a given information source $\widetilde{X}_n$, for $j = 1, \ldots, 8$. Compute $\mathcal{P}_8 = \mathcal{P}_8(\widetilde{X}_n)$ for each of the following $\widetilde{X}_n$.

(a) $\widetilde{X}_{16} = \{a_8, a_6, a_4, a_2, a_2, a_4, a_1, a_5, a_3, a_7, a_2, a_8, a_6, a_8, a_8, a_2\}$.
(b) $\widetilde{X}_{24} = \{a_1, a_1, a_1, a_7, a_6, a_5, a_7, a_5, a_4, a_8, a_4, a_8, a_2, a_2, a_3, a_1, a_2, a_3, a_7, a_5,$
    $a_6, a_7, a_1, a_8\}$.
(c) $\widetilde{X}_{32} = \{a_3, a_7, a_7, a_7, a_3, a_1, a_5, a_5, a_3, a_1, a_8, a_6, a_5, a_4, a_2, a_2, a_1, a_3, a_4, a_7,$
    $a_3, a_5, a_3, a_1, a_1, a_1, a_3, a_7, a_7, a_8, a_6, a_1\}$.

Exercise 4 As a continuation of Exercise 3, compute the entropy of each discrete probability distribution $\mathcal{P}_8 = \mathcal{P}_8(\widetilde{X}_n)$ for $\widetilde{X}_n = \widetilde{X}_{16}, \widetilde{X}_{24}, \widetilde{X}_{32}$.

Exercise 5 For the discrete probability distributions $\mathcal{P}_5 = \mathcal{P}_5(X_n)$ in Exercises 1–2, let $a_1 = p_1 + p_2$, $a_2 = p_3 + p_4 + p_5$, and consider the discrete probability distributions $A = \{a_1, a_2\}$, $Q_1 = \{p_1/a_1, \ p_2/a_1\}$, and $Q_2 = \{p_3/a_2, \ p_4/a_2, \ p_5/a_2\}$. Compute the entropies of A, Q_1, Q_2 for each of the information sources $X_n = X_8, X_{14}, X_{16}$ in Exercise 1a–c. Then apply the formula (5.1.19) of Theorem 3 to compute $H(\mathcal{P}_8) = H(\mathcal{P}_8(X_n))$. Compare with the answers for Exercise 2.

Exercise 6 For the discrete probability distributions $\mathcal{P}_8 = \mathcal{P}_8(\widetilde{X}_n)$ in Exercises 3–4, let $a_1 = p_1 + \cdots + p_4$, $a_2 = p_5 + \cdots + p_8$, and consider the probability distributions $\widetilde{A} = \{a_1, a_2\}$, $\widetilde{Q}_1 = \{p_1/a_1, \ldots, p_4/a_1\}$, and $\widetilde{Q}_2 = \{p_5/a_2, \ldots, p_8/a_2\}$. Compute the entropies $H(\widetilde{A}), H(\widetilde{Q}_1), H(\widetilde{Q}_2)$ for each of the information sources $\widetilde{X}_n = \widetilde{X}_{16}, \widetilde{X}_{24}, \widetilde{X}_{32}$ in Exercise 3a–c. Then apply (5.1.19) of Theorem 3 to compute $H(\mathcal{P}_8) = H(\mathcal{P}_8(\widetilde{X}_n))$. Compare with the answers for Exercise 4.

Exercise 7 Let $\widetilde{\mathcal{P}}_\ell$ for $\ell = 0, \ldots, n$ be defined by (5.1.18). Show that they are discrete probability distributions.

5.2 Binary Codes

In this section, we introduce the concept of binary codes and the Huffman coding scheme. As mentioned earlier, an encoded file or bit-stream is a "sequence" of numbers consisting of only zeros (0) and ones (1), but without separation of any punctuation such as commas, periods, hyphens, parentheses, quotation marks, etc. A typical code looks like

$$01011001100111011111000110001,$$

but is much much longer. By using a "code-table" that consists of "code-words", the message (such as a digital image, an audio recording, a document, etc.) can be recovered by "decoding" the bit-stream. A code-word is a binary number such as 0, 10, 110, 1111. If a table consists only of these four code-words; that is,

$$K_4 = \{0, 10, 110, 1111\}, \tag{5.2.1}$$

then to encode a sentence by using the table K_4 in (5.2.1), we can only use at most 4 different words in the sentence. For example, by using K_4 to represent the four words:

$$
\begin{aligned}
0 \quad &= \text{it} \\
10 \quad &= \text{good} \\
1111 &= \text{bad} \\
110 \quad &= \text{is}
\end{aligned}
$$

then the four bit-streams "110010", "01101111", "011010", "101101111" uniquely represent four different messages, without ambiguity. Try to decode (or decipher) them by yourself. (The answers are : "is it good", "it is bad", "it is good", and "good is bad".) Again, there is absolutely no ambiguity to use the code-table (5.2.1) to represent 4 different words to compose any sentence. Hence, the table K_4 is called a code-table. On the other hand, the table

$$T_4 = \{0, 10, 1, 11\}$$

is not a code-table. Indeed, by using T_4 to represent the four words:

$$
\begin{aligned}
0 \quad &= \text{you} \\
10 &= \text{not} \\
1 \quad &= \text{should} \\
11 &= \text{go}
\end{aligned}
$$

then the same bit-stream "011011" represents the following five different messages: "you should not go", "you go you go", "you should should you go", "you should should you should should", and "you go you should should".

Moreover, even among code-tables of the same size, some are better than others. For example, both

$$K_2 = \{0, 10\} \quad \text{and} \quad \widetilde{K}_2 = \{0, 01\}$$

are code-tables. By applying Table K_2 (and respectively Table $\widetilde{K}_2$) to represent the two colors b (black) and w (white); namely,

$$
\begin{aligned}
\text{Table } K_2 \quad & 0 = w, \quad 10 = b, \\
\text{Table } \widetilde{K}_2 \quad & 0 = w, \quad 01 = b,
\end{aligned}
$$

the two bit-streams

$$0001000101000 \tag{5.2.2}$$

$$0000100010100 \tag{5.2.3}$$

(with Table K_2 for decoding (5.2.2), and Table $\widetilde{K}_2$ for decoding (5.2.3)), represent the same color sequence

$$w \ w \ w \ b \ w \ w \ b \ b \ w \ w. \tag{5.2.4}$$

But there is some difference between them. To decode (5.2.2) by using Table K_2 to yield (5.2.4), it is "instantaneous" to identify the 3 black (b) pixels in (5.2.4), in that when reading from left to right the first 1 already determines the code-word 10 (and hence the color b), since the only code-word containing 1 is 10. On the other hand, to decode (5.2.3) by using Table $\widetilde{K}_2$ and again reading from left to right, identifying the color b is not instantaneous, since 0 does not indicate the code 01 without reading the second digit 1 in 01.

In any case, a binary code-word is always represented as a "sequence" of 0 and/or 1 without using "commas" in-between.

Definition 1 | **Length of code-word** | *The length of a code-word c_j in some code-table $K_N = \{c_1, \ldots, c_N\}$ is the number of digits 0 and/or 1 in the binary representation of the "integer" c_j, and will be denoted by*

$$\ell_j = \text{length } (c_j). \tag{5.2.5}$$

In constructing a code-table K_N with N code-words $c_1, \ldots, c_N$, it is of course most desirable to achieve the shortest lengths of the code-words. However, the restriction on the lengths is governed by the following **Kraft's inequality**:

$$\sum_{j=1}^{N} 2^{-\ell_j} \leq 1. \tag{5.2.6}$$

Observe that if $x_1 = 0$ and $x_2 = 1$ are both in the code-table K_N, then there cannot be any other code-word in K_N, since we would already have attained the upper bound in (5.2.6); namely

$$2^{-\ell_1} + 2^{-\ell_2} = \frac{1}{2} + \frac{1}{2} = 1,$$

so that the size of K_N is $N = 2$. We will not prove (5.2.6) in this book, but mention that Kraft only proved (5.2.6) for instantaneous codes in his 1949 MIT Master's Thesis in Electrical Engineering, but McMillan removed the "instantaneous" restriction to allow the validity of (5.2.6) for all decodable (or decipherable) tables.

Remark 1 To assess the suitability of the code-tables $K_N = \{c_1, \ldots, c_N\}$ in Definition 1 for encoding a given information source $X_n = \{x_1, \ldots, x_n\}$ with a relatively shorter bit-stream (or smaller file size), recall the notion of the discrete probability distribution $\mathcal{P} = \{p_1, \ldots, p_m\}$ associated with X_n. Assuming that the code-table K_N is constructed solely for X_n with probability distribution $\mathcal{P}$, then since each

code-word $c_j \in K_N$ is constructed according to the value of the corresponding $p_j \in \mathcal{P}$, as compared with all $p_k \in \mathcal{P}, k = 1, \ldots, m$, the cardinality $m = \#\mathcal{P}$ agrees with the number N of code-words in K_N. Under this assumption, observe that the precise value of $p_j \in \mathcal{P}$, for each fixed j, as defined by the histogram of X_n in (5.1.2), increases proportionally as the frequency of occurrence of the same $x_j \in X$ according to (5.1.2) increases. Hence, for this discussion with $m = N$, the length ℓ_j of the code-word $c_j \in K_N$ should be relatively shorter for larger relative values of p_j for the code-table K_N to be suitable for encoding X_n. For this reason, the values $p_1, \ldots, p_m \in \mathcal{P}$ are used as weights to define the weighted average:

$$\text{avglength}\,(K_N) = \text{avglength}\,\{c_1, \ldots, c_N\} = \sum_{j=1}^{N} p_j \ell_j, \qquad (5.2.7)$$

called "**average code-word length**" of K_N, where $\ell_j = \text{length}\,(c_j)$ as introduced in (5.2.5). ∎

Remark 2 In applications, as mentioned in Remark 1, the same probability distribution $\mathcal{P}$, and hence the same code-table K_N, is constructed for a large class of information sources X_n with different cardinalities $n = \#X_n$. In particular, N is usually larger than $m = \#\mathcal{P}$, where $\mathcal{P}$ is the discrete probability distribution associated with X_n. For example, the same code-table K_N, called the "Huffman table", to be discussed later in this section, is used for most (if not all) JPEG compressed digital images, a topic of investigation in Sect. 5.4 of this chapter. ∎

Let us return to Kraft's inequality (5.2.6) and observe that it governs the necessity of long code-word lengths, when the number of code-words must be sufficiently large for practical applications. However, it does not give a quantitative measurement of the code-word lengths. In the following, we will see that the entropy $H(\mathcal{P})$ of a given discrete probability distribution $\mathcal{P} = \{p_1, \ldots, p_N\}$, with N equal to the size of the desired code-table $K_N = \{c_1, \ldots, c_N\}$, provides a perfect measurement stick.

Theorem 1 *Let $K_N = \{c_1, \ldots, c_N\}$ be a code-table with code-word lengths $\ell_j = \text{length}\,\{c_j\}, j = 1, \ldots, N$, as introduced in Definition 1. Then the average code-word length of K_N is bounded below by the entropy of the desired discrete probability distribution that governs the construction of the code-table; namely,*

$$H(\mathcal{P}) \leq \text{avglength}\,(K_N), \qquad (5.2.8)$$

where avglength (K_N) is defined in (5.2.7). Furthermore, equality in (5.2.8) is achieved if and only if both of the following conditions are satisfied:

$$p_j = \frac{1}{2^{n_j}}, \quad j = 1, \ldots, N, \qquad (5.2.9)$$

for some positive integers $n_1, \ldots, n_N$; and

$$\ell_j = n_j, \quad j = 1, \ldots, N, \tag{5.2.10}$$

so that Kraft's inequality becomes equality.

Proof The proof of this theorem is an application of Theorem 1 and Kraft's inequality (5.2.6). Indeed, let

$$q_j = \frac{1}{C} \, 2^{-\ell_j}, \quad j = 1, \ldots, N,$$

where

$$C = \sum_{k=1}^{N} 2^{-\ell_k}.$$

Then since $q_1 + \ldots + q_N = 1$ and $0 \le q_j \le 1$ for all j, we may apply (5.1.5) on p.212 to conclude that

$$
\begin{aligned}
H(\mathcal{P}) \le G(q_1, \ldots, q_N) &= \sum_{j=1}^{N} p_j \log_2 \left(C \, 2^{\ell_j} \right) \\
&= \sum_{j=1}^{N} p_j \left(\ell_j + \log_2 C \right) \\
&= \text{avglength} \, (K_N) + \log_2 C,
\end{aligned}
$$

according to (5.2.7) and $\sum_{j=1}^{N} p_j = 1$. In view of Kraft's inequality, we have $\log_2 C \le 0$, and hence we have completed the derivation of (5.2.8).

Furthermore, equality in (5.2.8) holds if and only if $\log_2 C = 0$, or $C = 1$ and

$$H(\mathcal{P}) = G(q_1, \ldots, q_n),$$

which, according to the last statement in Theorem 1 of the previous section on p.211, is equivalent to

$$q_1 = p_1, \ldots, q_N = p_N,$$

or equivalently,

$$p_1 = \frac{1}{2^{\ell_1}}, \ldots, p_N = \frac{1}{2^{\ell_N}}.$$

This completes the proof of (5.2.9)–(5.2.10), with $n_j = \ell_j$, which is an integer, being the length of the code-word c_j, $j = 1, \ldots, N$. ∎

Remark 3 Since discrete probability distributions are seldom positive integer powers of $\frac{1}{2}$, one cannot expect to achieve a code-table with minimum average code-word lengths in general. On the other hand, there are fairly complicated coding schemes,

such as "arithmetic coding", that could reach the entropy lower bound as close as desired.   ∎

In introducing the notion of entropy, Claude Shannon also showed that the entropy can be used as a measuring stick in that there exist instantaneous code-tables with average code-word lengths bounded above by the entropy plus 1. In other words, we have the following result, called "noiseless coding" by Shannon.

Theorem 2 | **Noiseless Coding Theorem** | *For any discrete probability distribution* $\mathcal{P} = \{p_1, \ldots, p_N\}$,

$$H(\mathcal{P}) \leq \min_{K_N} \text{ avglength } (K_N) < H(\mathcal{P}) + 1,$$

where the minimum is taken over all instantaneous code-tables.

We will not prove the above theorem in this book, but proceed to discuss the most popular coding scheme introduced by David Huffman in his 1952 paper.

In Professor Fano's undergraduate class at MIT in 1951, in lieu of taking the final examination, the students were given the option of turning in a project paper on developing a coding scheme that achieves entropy bounds in the Noiseless Coding Theorem. One of the students, David Huffman, met the challenge and developed the following simple, yet effective, scheme based on building a binary tree by using the given probability distribution $\mathcal{P}$. We remark that since only $\mathcal{P}$ and not the information source is used to construct the code-table, there is no need to refer to the information source X_n as in Definition 1.

Algorithm 1 | **Huffman coding scheme** | Let $\mathcal{P}_N^0 = \{p_1, \ldots, p_N\}$ be a given discrete probability distribution. For convenience, assume that $p_1, \ldots, p_N$ have been re-arranged so that
$$0 < p_1 \leq p_2 \leq \cdots \leq p_N.$$

Step 1. Use the two smallest numbers p_1, p_2 to build the first branch of the Huffman tree, with labels p_1 and p_2 for the two branches and $p_1 + p_2$ for the "stem" of the branch. Then assign the partial codes 0 to the smaller p_1 and 1 to the larger p_2, as follows. (If $p_1 = p_2$, the assignment is arbitrary.)
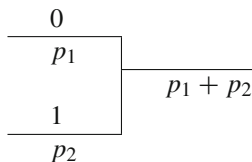
Step 2. Let $\mathcal{P}_{N-1}^1 = \{p_1^1, \ldots, p_{N-1}^1\}$ be the discrete probability distribution consisting of
$$p_1 + p_2, p_3, \ldots, p_N,$$

but re-arranged in non-decreasing order, so that

$$0 < p_1^1 \leq p_2^1 \leq \cdots \leq p_{N-1}^1.$$

(i) If $p_1 + p_2 \neq p_1^1, p_2^1$, repeat Step 1 to build a new branch with analogous labels and partial codes as follows.

$$
\begin{array}{l}
0 \\
\overline{} \\
p_1^1 = p_3 \\
 p_1^1 + p_2^1 = p_3 + p_4 \\
1 \\
p_2^1 = p_4
\end{array}
$$

(ii) If $p_1 + p_2 = p_1^1$, then the same labeling along with partial code assignments yields the following branch.

$$
\begin{array}{l}
0 \\
p_1 \\
 0 \\
 p_1 + p_2 = p_1^1 \\
1 \\
p_2 \\
 p_1^1 + p_2^1 = p_1 + p_2 + p_3 \\
 1 \\
 p_2^1 = p_3
\end{array}
$$

(iii) Similarly, if $p_1 + p_2 = p_2^1$, then the same labeling along with partial code assignments yields the following branch.

$$
\begin{array}{l}
 0 \\
 p_1^1 = p_3 \\
0 p_1^1 + p_2^1 = p_1 + p_2 + p_3 \\
p_1 1 \\
 p_1 + p_2 = p_2^1 \\
1 \\
p_2
\end{array}
$$

Step 3. Let $\mathcal{P}_{N-2}^2 = \{p_1^2, \ldots, p_{N-2}^2\}$ be the discrete probability distribution consisting of

$$p_1^1 + p_2^1, \ p_3^1, \ \cdots, \ p_{N-1}^1,$$

but re-arranged in non-decreasing order, so that

$$0 < p_1^2 \le p_2^2 \le \cdots \le p_{N-2}^2.$$

Repeat Step 2 to build the following sub-branch and assign partial codes as follows.

$$
\begin{array}{ll}
0 & \\
p_1^2 & \\
& p_1^2 + p_2^2 \\
1 & \\
p_2^2 &
\end{array}
$$
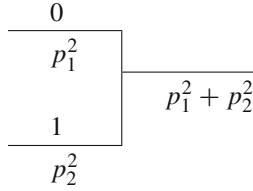
(i) If $p_1^2 = p_1^1 + p_2^1 = p_3 + p_4$ as in Step 2 (i), the full sub-branch becomes

$$
\begin{array}{ll}
0 & \\
p_3 & \\
& 0 \\
& p_3 + p_4 \\
1 & \\
p_4 & \\
& 1 \\
& p_2^2
\end{array}
$$

for $p_3 + p_4 \le p_2^2$; or

$$
\begin{array}{ll}
& 0 \\
& p_2^1 \\
0 & \\
p_3 & \\
& 1 \\
& p_3 + p_4 \\
1 & \\
p_4 &
\end{array}
$$
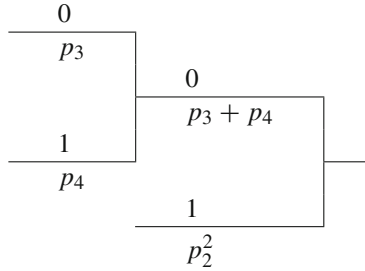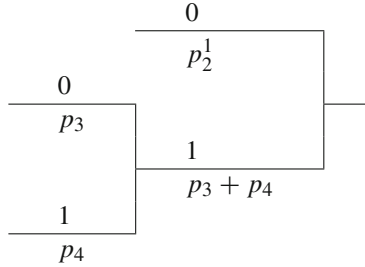
for $p_3 + p_4 \ge p_2^1$.

(ii) If $p_1^2 \ne p_3 + p_4$, then by extending the sub-branches in Step 2 (ii) or (iii), we have one of the following full sub-branches.

Step 4. Repeat the above steps till we arrive at $\mathcal{P}_2^{N-2} = \{p_1^{N-2}, p_2^{N-2}\}$ with re-arrangement: $p_1^{N-2} \leq p_2^{N-2}$. Then this final partial sub-branch has the trunk of the full tree as its stem, as follows.

Step 5. Formulate the code-word corresponding to each p_j, for $j = 1, \ldots, N$, by tracing from the branch that starts with the label p_j all the way to the trunk (with label $p_1 + \cdots + p_N = 1$), by putting the partial codes 0 or 1 on each branch together, but in **reverse order**. For example, the code-word corresponding to p_j in the following tree is 01101 (not 10110).

Remark 4 The re-arrangement in non-decreasing order in $\mathcal{P}_N^0$ and each step of $\mathcal{P}_{N-1}^1, \ldots, \mathcal{P}_2^{N-2}$ is not necessary, but is stated in Algorithm 1 for convenience to describe the coding scheme. Furthermore, the code-table (called the Huffman table) corresponding to $\mathcal{P} = \mathcal{P}_N^0$ is not unique, since some of the p_j's and/or p_j^ℓ's could be the same.   ∎

Example 1 Let $\mathcal{P} = \left\{\frac{1}{24}, \frac{6}{24}, \frac{2}{24}, \frac{4}{24}, \frac{1}{24}, \frac{2}{24}, \frac{8}{24}\right\}$. Follow the steps in the Huffman coding scheme of Algorithm 1 to construct a Huffman table. Verify that the code-table satisfies both the equality in Kraft's inequality and the Noiseless Coding entropy bounds.

Solution Re-arrange the probability distributions in non-decreasing order. That is, set

$$\mathcal{P} = \mathcal{P}_7^0 = \left\{ \frac{1}{24}, \frac{1}{24}, \frac{2}{24}, \frac{2}{24}, \frac{4}{24}, \frac{6}{24}, \frac{8}{24} \right\} = \{p_1, \ldots, p_7\}.$$

Step 1.
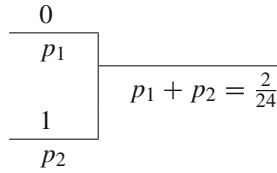
Step 2.

$$\mathcal{P}_6^1 = \{p_1 + p_2, p_3, \ldots, p_7\} = \{p_1^1, p_2^1, \ldots, p_6^1\}$$
$$= \left\{ \frac{2}{24}, \frac{2}{24}, \frac{2}{24}, \frac{4}{24}, \frac{6}{24}, \frac{8}{24} \right\}$$
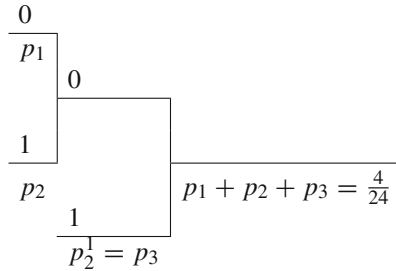
Step 3.

$$\mathcal{P}_5^2 = \{p_1^2, p_2^2, p_3^2, p_4^2, p_5^2\} = \{p_4, p_1 + p_2 + p_3, p_5, p_6, p_7\}$$
$$= \left\{ \frac{2}{24}, \frac{4}{24}, \frac{4}{24}, \frac{6}{24}, \frac{8}{24} \right\}$$
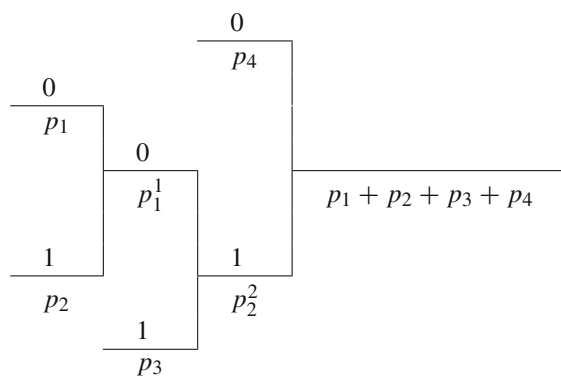
Step 4.

$$\mathcal{P}_4^3 = \{p_1^3, p_2^3, p_3^3, p_4^3\} = \{p_5, p_1 + \ldots + p_4, p_6, p_7\}$$
$$= \left\{ \frac{4}{24}, \frac{6}{24}, \frac{6}{24}, \frac{8}{24} \right\}$$
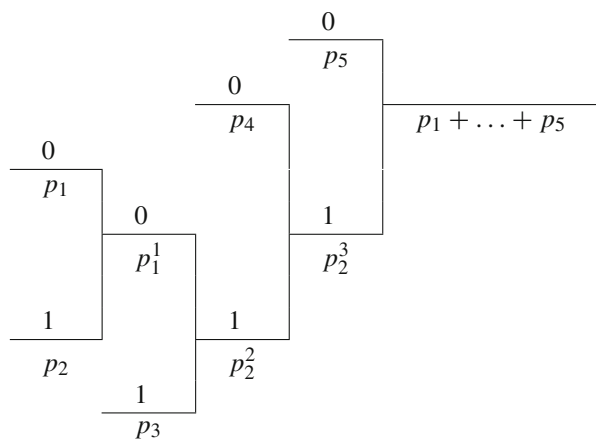
Step 5.

$$\mathcal{P}_3^4 = \{p_1^4, p_2^4, p_3^4\} = \{p_6, p_7, p_1 + p_2 + p_3 + p_4 + p_5\}$$
$$= \left\{\frac{6}{24}, \frac{8}{24}, \frac{10}{24}\right\}$$
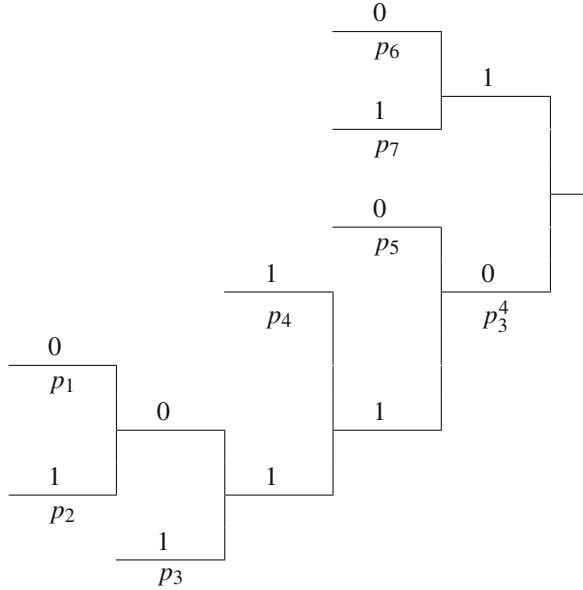
In conclusion, tracing the 0's and 1's in reverse order for each of $p_1, \ldots, p_7$, the code-table is given by

$$K_7 = \{01100, 01101, 0111, 010, 00, 10, 11\},$$

with code lengths given by

$$\ell_1 = 5, \ell_2 = 5, \ell_3 = 4, \ell_4 = 3, \ell_5 = 2, \ell_6 = 2, \ell_7 = 2.$$

Hence,

$$\sum_{j=1}^{n} 2^{-\ell_j} = \frac{1}{32} + \frac{1}{32} + \frac{1}{16} + \frac{1}{8} + \frac{1}{4} + \frac{1}{4} + \frac{1}{4} = 1,$$

which attains the equality in Kraft's inequality. Finally, the entropy of $\mathcal{P}$ is :

$$H(\mathcal{P}) = \sum_{j=1}^{7} p_j \log_2 \frac{1}{p_j} = \frac{1}{24} \log_2 24 + \frac{6}{24} \left(\log_2 24 - \log_2 6 \right)$$

$$+ \frac{2}{24} \left(\log_2 24 - \log_2 2 \right) + \frac{4}{24} \left(\log_2 24 - \log_2 4 \right) + \frac{1}{24} \log_2 24$$

$$+ \frac{2}{24} \left(\log_2 24 - \log_2 2 \right) + \frac{8}{24} \left(\log_2 24 - \log_2 8 \right)$$

$$= \log_2 24 - \frac{1}{24} \left(0 + 6(\log_2 6) + 2 + 4\log_2 4 + 0 + 2 + 8(\log_2 8) \right)$$

$$= 2.4387,$$

and the average code-word length is:

$$\text{avglength } (K_7) = \frac{1}{24} \cdot 5 + \frac{1}{24} \cdot 5 + \frac{2}{24} \cdot 4 + \frac{2}{24} \cdot 3 + \frac{4}{24} \cdot 2 + \frac{6}{24} \cdot 2 + \frac{8}{24} \cdot 2$$

$$= \frac{1}{24} (5 + 5 + 8 + 6 + 8 + 12 + 16) = \frac{60}{24} = 2.5.$$

Hence, $H(\mathcal{P}) = 2.4387 < 2.5 < 3.4387 = H(\mathcal{P}) + 1$, as required by the Noiseless Coding Theorem.

We remark that, since the probabilities $p_1, \ldots, p_7$, with the exception of $\frac{6}{24} = \frac{1}{4}$, are not integer powers of $\frac{1}{2}$, it follows from Theorem 1 that the average code-word length cannot attain the entropy lower bound of 2.4387, although 2.5 is much closer to the lower bound than the upper bound of $H(\mathcal{P}) + 1 = 3.4387$. ∎

<div style="text-align:center">

Exercises

</div>

Exercise 1 Let $f(x)$ be the function defined by $f(x) = x \log_2 x, 0 < x \leq 1$, and $f(0) = 0$. Show that $f(x)$ is continuous on the closed interval $[0, 1]$ and has continuous derivative for $0 < x \leq 1$.

Exercise 2 For the discrete probability distributions $\mathcal{P}_5 = \mathcal{P}_5(X_n)$ in Exercise 1 on p.216, apply Algorithm 1 to construct a Huffman table for each of $X_n = X_8, X_{14}, X_{16}$. Also, compute the average code-word length, avglength $(K_5(X_n))$, again for each of $X_n = X_8, X_{14}, X_{16}$. Finally, compare with the entropy bounds of the Noiseless Coding Theorem by using the results of Exercise 2 on p.216

Exercise 3 For the discrete probability distributions $\mathcal{P}_8 = \mathcal{P}_8(\widetilde{X}_n)$ in Exercise 3 on p.217, apply Algorithm 1 to construct a Huffman table for each of $\widetilde{X}_n = \widetilde{X}_{16}, \widetilde{X}_{24}, \widetilde{X}_{32}$. Also, compute the average code-word length, avglength $(K_8(\widetilde{X}_n))$, for each of $\widetilde{X}_n = \widetilde{X}_{16}, \widetilde{X}_{24}, \widetilde{X}_{32}$. Finally, compare with the entropy bounds of the Noiseless Coding Theorem by using the results of Exercise 4 on p.217.

Exercise 4 Show that the Huffman coding schemes achieve the lower bound for any probability distribution $\{p_1, \ldots, p_m\}$ with $p_j = 2^{-n_j}$, where n_j is a positive integer, for $j = 1, \ldots, m$.

Exercise 5 (For advanced students) Provide a proof of Kraft's inequality (5.2.6).

Exercise 6 (For advanced students) Provide a proof of the upper bound $H(\mathcal{P}) + 1$ in Theorem 2, the Noiseless Coding Theorem.

5.3 Lapped Transform and Compression Schemes

In this current era of "information revolution", data compression is a necessity rather than convenience or luxury. Without the rapid advancement of the state-of-the-art compression technology, not only the internet highway is unbearably over-crowded, but data management would be a big challenge also. In addition, data transmission would be most competitive and data storage would be very costly.

On the other hand, for various reasons, including strict regulations (such as storage of medical images), industry standards (for the capability of data retrieval by non-proprietary hardware and software installed in PC's or hand-held devices), and the need for a user-friendly environment, the more complex and/or proprietary solutions are not used by the general public, though they usually have better performance. For instance, ASCII is the preferred text format for word processors and Huffman coding is the most popular binary coding scheme, even though arithmetic coding is more optimal.

In general, there are two strategies in data compression, namely: lossless compression and lossy compression. Lossless compression is reversible, meaning that the restored data file is identical to the original data information. There are many applications that require lossless compression. Examples include compression of executable codes, word processing files, and to some extent, medical records, and medical images. On the other hand, since lossless compression does not satisfactorily reduce file size in general, lossy compression is most widely used.

Lossy compression is non-reversible, but allows insignificant loss of the original data source, in exchange for significantly smaller compressed file size. Typical applications are image, video, and audio compressions. For compression of digital images and videos, for instance, compressed imagery is often more visually pleasing than the imagery source, which inevitably consists of (additive) random noise due to perhaps insufficient lighting and non-existence of "perfect sensors" for image capture. Since random noise increases entropy, which in turn governs the lower bound of the compressed file size according to the Noiseless Coding Theorem (see Theorem 2 on p.222), lossy compression via removing a fair amount of such additive noise is a preferred approach. The topic of noise removal will be studied in a forthcoming publication of this book series, "Mathematics Textbooks for Science and Engineering (MTSE)".

There are three popular methods for lossless data compression: (i) run-length encoding (RLE, as briefly mentioned in the introduction of Sect. 3.3), (ii) delta encoding, and (iii) entropy coding. RLE simply involves encoding the number of the same source word that appears repeatedly in a row. For instance, to encode a one-bit line drawing along each scan-line, only very few dots of the line drawing are on the scan-line. Hence, if "1" is used for the background and "0" for the line drawing, there are long rows of repeating "1"'s before a "0" is encountered. Therefore, important applications of RLE include graphic images (cartoons) and animation movies. It is also used to compress Windows 3.x bitmap for the computer startup screen. In addition, RLE is incorporated with Huffman encoding for the JPEG compression standard to be discussed in some details later in Sect. 5.4

Delta encoding is a simple idea for encoding data in the form of differences. In Example 1 on p.210, the notion of "differential pulse code modulation (DPCM)" is introduced to create long rows of the same source word for applying RLE. In general, delta encoding is used only as an additional coding step to further reduce the encoded file size. In video compression, "delta frames" can be used to reduce frame size and is therefore used in every video compression standard. We will also mention DPCM in JPEG compression in the next section.

Perhaps the most powerful stand-alone lossless compression scheme is LZW compression, named after the developers A. Lempel, J. Ziv, and T. Welch. Typical applications are compression of executable codes, source codes, tabulated numbers, and data files with extreme redundancy. In addition, LZW is used in GIF image files and as an option in TIFF and PostScript. However, LZW is a proprietary encoding scheme owned by Unisys Corporation.

Let us now focus on the topic of lossy compression, to be applied to the compression of digital images and video, the topic of discussion in the next section. The general overall lossy compression scheme consists of three steps:

  (i) Transformation of source data,
 (ii) Quantization,
(iii) Entropy coding.

Both (i) and (iii) are reversible at least in theory, but (ii) is not. To recover the information source, the three steps are:

  (i) De-coding by applying the code-table,
 (ii) De-quantization,
(iii) Inverse transformation.

We have briefly discussed, in Example 2 of Sect. 5.1, that an appropriate transformation could be introduced to significantly reduce the entropy of certain source data, without loss of any data information. This type of specific transformations is totally data-dependent and therefore not very practical. Fortunately, there are many transformations that can be applied for sorting data information without specific knowledge of the data content, though their sole purpose is not for reduction of the data entropy. When data information is properly sorted out, the less significant content can be suppressed and the most insignificant content can be eliminated, if desired, so as to

reduce the entropy. As a result, shortened binary codes, as governed by the Noiseless Coding Theorem studied in the previous section, can be constructed.

The most common insignificant data content is (additive) random noise, with probability values densely distributed on the (open) unit interval $(0, 1)$. Hence, embedded with such noise, the source data has undesirably large entropy. Fortunately, partially due to the dense distribution, random noise lives in the high-frequency range. Therefore, all transformations that have the capability of extracting high-frequency contents could facilitate suppressing the noise content by means of "quantization", to be discussed in the next paragraph. Such transformations include DCT and DFT studied in Chap. 4, DST (discrete sine transform), Hardamard transform, and DWT (discrete wavelet transform), which will be introduced and studied in some depth in Chaps. 8 and 9. Among all of these transformations, DCT, particularly DCT-II, and DCT-IV introduced in Sect. 4.4 of Chap. 4, is the most popular for compression of digital images, videos, and digitized music.

The key to the feasibility of significant file size reduction for lossy compression is the "quantization" process that maps a "fine" set of real numbers to a "coarse" set of integers. Of course such mappings are irreversible. For any real numbers x, $sgn\, x$ (called the sign of x) is defined by

$$sgn\, x = \begin{cases} 1, & \text{for } x > 0, \\ 0, & \text{for } x = 0, \\ -1, & \text{for } x < 0. \end{cases}$$

Also, for any non-negative real number x, $\lfloor x \rfloor$ will denote the largest integer not exceeding x. Then the most basic quantization process is the mapping of $x \in \mathbb{R}$ to an integer $\widetilde{x}$, defined by

$$\widetilde{x} = \text{round}\left(\frac{x}{Q}\right) = (sgn\, x)\left\lfloor \frac{|x|}{Q} \right\rfloor, \tag{5.3.1}$$

where Q is a positive integer, called the "quantizer" of the "round-off" function defined in (5.3.1). Hence, $\widetilde{x}$ is an integer and $Q\widetilde{x}$ is an approximation of x, in that

$$|x - Q\widetilde{x}| < 1.$$

A better approximation of the given real number x could be achieved by $Q\widehat{x}$, with $\widehat{x}$ defined by

$$\widehat{x} = (sgn\, x)\left\lfloor \frac{|x \pm \lfloor \frac{Q}{2} \rfloor|}{Q} \right\rfloor, \tag{5.3.2}$$

where the "+" sign or "−" sign is determined by whichever choice yields the smaller $|x - Q\widehat{x}|$. In any case, it is clear that the binary representation of $\widetilde{x}$ (or of $\widehat{x}$) requires fewer bits for larger integer values of the quantizer Q. In applications to audio and image compressions, since the human ear and human eye are less sensitive to higher

frequencies, larger values of the quantizer Q can be applied to higher-frequency DCT terms to save more bits. Moreover, since additive random noise lives in the higher frequency range, such noise could be suppressed, often resulting in more pleasing audio and imagery quality.

In summary, lossy compression is achieved by binary encoding (such as Huffman coding) of the quantized values of the transformed data; and recovery of the source data is accomplished by applying the inverse transformation to the de-quantized values of the decoded data. Since the quantization process is irreversible, the compression scheme is lossy.

As mentioned above, DCT is the most popular transformation, particularly for audio, image, and video compressions. However, for such large data sets, it is not feasible to apply DCT to the entire data source. The common approach is to partition the data set into segments (also called tiles or blocks) and apply DCT to each segment. Unfortunately, quantization of the DCT terms for individual segments introduces "*blocky artifacts*", called "*quantization noise*". To suppress such "*noise*", an appropriate pre-processing scheme can be incorporated with the discrete cosine transformation for each segment. We will only consider n-point DCT, and in particular, DCT-I, DCT-II, DCT-III, and DCT-IV, as studied in Sect. 4.4 of Chap. 4. Since quantization noise occurs only at or near the boundaries (i.e. end-points) of the partitioned data segments, data pre-processing should be applied only to the boundaries while leaving the interior data content unaltered. More precisely, let us consider any (finite) data sequence

$$U = \{u_0, \ldots, u_{N-1}\}$$

of real numbers. For convenience, we assume that N is divisible by the length n of the segments

$$U_\ell = \{u_{\ell n}, \ldots, u_{(\ell+1)n-1}\}, \ \ell = 0, \ldots, L-1, \tag{5.3.3}$$

where $n > 0$. In other words, we have $N = Ln$ and

$$U = \{U_0, \ldots, U_{L-1}\}. \tag{5.3.4}$$

Consider any desired positive even integer $m \leq n$, and two appropriate pre-processing filters

$$H_1 = \left\{ h_{-\frac{m}{2}}, \ldots, h_{\frac{m}{2}-1} \right\}; \ H_3 = \left\{ h_{n-\frac{m}{2}}, \ldots, h_{n+\frac{m}{2}-1} \right\} \tag{5.3.5}$$

of real numbers for the purpose of pre-processing the data segments U_ℓ in (5.3.3), with the first filter applied to the left boundary and the second filter applied to the right boundary, respectively. Depending on the choice of the DCT, the two filters must be so chosen that the original data segments $U_\ell, \ell = 0, \ldots, L-1$, in (5.3.3) can be recovered from the DCT of the pre-processed data.

Example 1 Let $C_n = [c_0, \ldots, c_{n-1}]$ denote the n-point DCT-IV introduced in Sect. 4.4 of Chap. 4, with column vectors

$$\mathbf{c}_k = \left[\sqrt{\frac{2}{n}} \cos \frac{\frac{1}{2}(k+\frac{1}{2})\pi}{n}, \sqrt{\frac{2}{n}} \cos \frac{\frac{3}{2}(k+\frac{1}{2})\pi}{n}, \ldots, \sqrt{\frac{2}{n}} \cos \frac{(n-\frac{1}{2})(k+\frac{1}{2})\pi}{n} \right]^T$$

for $k = 0, \ldots, n-1$. Construct a pre-processing scheme so that the original data can be recovered via the inverse DCT, $C_n^{-1} = C_n^T$, from the DCT of the pre-processed data.

Solution For the interior segments U_ℓ, $\ell = 1, \ldots, L-2$, let

$$\widetilde{U}_\ell = \{\widetilde{u}_{\ell n}, \ldots, \widetilde{u}_{(\ell+1)n-1}\}$$

be defined as follows:

$$\begin{cases} \widetilde{u}_{\ell n} = h_0 u_{\ell n} + h_{-1} u_{\ell n - 1}, \\ \quad\quad\vdots \\ \widetilde{u}_{\ell n + \frac{m}{2} - 1} = h_{\frac{m}{2}-1} u_{\ell n + \frac{m}{2} - 1} + h_{-\frac{m}{2}} u_{\ell n - \frac{m}{2}}, \end{cases} \tag{5.3.6}$$

$$\begin{cases} \widetilde{u}_{\ell n + \frac{m}{2}} = u_{\ell n + \frac{m}{2}}, \\ \quad\quad\vdots \\ \widetilde{u}_{(\ell+1)n - \frac{m}{2} - 1} = u_{(\ell+1)n - \frac{m}{2} - 1}, \end{cases} \tag{5.3.7}$$

$$\begin{cases} \widetilde{u}_{(\ell+1)n - \frac{m}{2}} = h_{n - \frac{m}{2}} u_{(\ell+1)n - \frac{m}{2}} - h_{n + \frac{m}{2} - 1} u_{(\ell+1)n + \frac{m}{2} - 1}, \\ \quad\quad\vdots \\ \widetilde{u}_{(\ell+1)n - 1} = h_{n-1} u_{(\ell+1)n-1} - h_n u_{(\ell+1)n}, \end{cases} \tag{5.3.8}$$

for $\ell = 1, \ldots, L-2$. For the first segment U_0, let

$$\widetilde{U}_0 = \{u_0, \ldots, u_{n-\frac{m}{2}-1}, \widetilde{u}_{n-\frac{m}{2}}, \ldots, \widetilde{u}_{n-1}\},$$

where $\widetilde{u}_{n-\frac{m}{2}}, \ldots, \widetilde{u}_{n-1}$ are defined by (5.3.8) with $\ell = 0$, and for the last segment U_{L-1}, let

$$\widetilde{U}_{L-1} = \{\widetilde{u}_{(L-1)n}, \ldots, \widetilde{u}_{(L-1)n+\frac{m}{2}-1}, u_{(L-1)n+\frac{m}{2}}, \ldots, u_{Ln-1}\},$$

where $\widetilde{u}_{(L-1)n}, \ldots, \widetilde{u}_{(L-1)n+\frac{m}{2}-1}$ are defined by (5.3.6) with $\ell = L-1$. To recover the original data $U = \{U_1, \ldots, U_{L-1}\}$ in (5.3.4) from the DCT content

$$C_n \widetilde{U}_\ell, \ \ell = 0, \ldots, L-1,$$

where C_n is the n-point DCT-IV matrix and the sequence $\widetilde{U}_\ell$ is also considered as a column vector, namely:

$$\widetilde{U}_\ell = \left[\widetilde{u}_{\ell n}, \dots, \widetilde{u}_{(\ell+1)n-1} \right]^T,$$

we proceed as follows. Write $C_n \widetilde{U}_\ell = \left[v_{\ell n}, \dots, v_{(\ell+1)n-1} \right]^T$ for $\ell = 0, \dots, L-1$, and consider the DCT data sequence

$$V = \{v_0, \dots, v_{Ln-1}\} = \{v_0, \dots, v_{N-1}\}, \tag{5.3.9}$$

partitioned into the following L disjoint segments of length n:

$$V_\ell = \{v_{\ell n}, \dots, v_{(\ell+1)n-1}\}, \ \ell = 0, 1, \dots, L-1. \tag{5.3.10}$$

In the following, we consider these sequences as column vectors.

Next, we extend the inverse $C_n^{-1} = C_n^T$ to an $(n+m) \times n$ matrix B by tacking on $\frac{m}{2}$ rows, $\mathbf{c}_{-\frac{m}{2}}^T, \dots, \mathbf{c}_{-1}^T$ to the top; and $\frac{m}{2}$ rows, $\mathbf{c}_n^T, \dots, \mathbf{c}_{n+\frac{m}{2}-1}^T$, to the bottom, where $\mathbf{c}_k$ denotes the same kth column of the n-point DCT-IV matrix, but extended beyond $0 \leq k \leq n-1$. Precisely, the transpose B^T of the $(n+m) \times n$ matrix B is given by

$$B^T = \left[\mathbf{c}_{-\frac{m}{2}}, \dots, \mathbf{c}_{-1}, C_n, \mathbf{c}_n, \dots, \mathbf{c}_{n+\frac{m}{2}-1} \right]$$
$$= \left[\mathbf{c}_{-\frac{m}{2}}, \dots, \mathbf{c}_{n+\frac{m}{2}-1} \right].$$

We will apply B to each $V_\ell, \ell = 1, \dots, L-2$, and again consider V_ℓ as a column vector. For $\ell = 0$ and $\ell = L-1$, we only need

$$B_0^T = \left[C_n, \mathbf{c}_n, \dots, \mathbf{c}_{n+\frac{m}{2}-1} \right],$$
$$B_1^T = \left[\mathbf{c}_{-\frac{m}{2}}, \dots, \mathbf{c}_{-1}, C_n \right], \tag{5.3.11}$$

respectively. From the definition of V and V_ℓ in (5.3.9)–(5.3.10), we have, for $1 \leq \ell \leq L-2$,

$$V_\ell = \left[\mathbf{c}_{-\frac{m}{2}}, \dots, \mathbf{c}_{-1}, C_n, \mathbf{c}_n, \dots, \mathbf{c}_{n+\frac{m}{2}-1} \right] \begin{bmatrix} \widetilde{u}_{\ell n - \frac{m}{2}} \\ \vdots \\ \widetilde{u}_{(\ell+1)n+\frac{m}{2}-1} \end{bmatrix}$$
$$= B^T \left[\widetilde{u}_{\ell n - \frac{m}{2}}, \dots, \widetilde{u}_{\ell n - 1}, \widetilde{u}_{\ell n}, \dots, \widetilde{u}_{(\ell+1)n-1}, \widetilde{u}_{(\ell+1)n}, \dots, \widetilde{u}_{(\ell+1)n+\frac{m}{2}-1} \right]^T.$$

Hence, when B is applied to V_ℓ for $1 \leq \ell \leq L-2$, B_0 to V_0, and B_1 to $V_{\ell-1}$, we obtain

$$\left[w_0^0, \dots, w_{n+\frac{m}{2}-1}^0 \right]^T = B_0 B_0^T \left[u_0, \dots, u_{n-\frac{m}{2}-1}, \widetilde{u}_{n-\frac{m}{2}}, \dots, \widetilde{u}_{n+\frac{m}{2}-1} \right]^T, \tag{5.3.12}$$

$$\left[w^{L-1}_{(L-1)n-\frac{m}{2}}, \ldots, w^{L-1}_{Ln-1}\right]^T = B_1 B_1^T \left[\tilde{u}_{(L-1)n-\frac{m}{2}}, \ldots, \tilde{u}_{(L-1)n+\frac{m}{2}-1},\right.$$

$$\left. u_{(L-1)n+\frac{m}{2}}, \ldots, u_{Ln-1}\right]^T,$$

$$\left[w^{\ell}_{\ell n-\frac{m}{2}}, \ldots, w^{\ell}_{(\ell+1)n+\frac{m}{2}-1}\right]^T = BB^T \left[\tilde{u}_{\ell n-\frac{m}{2}}, \ldots, \tilde{u}_{(\ell+1)n+\frac{m}{2}-1}\right]^T,$$

for $\ell = 1, \ldots, L - 2$.

For DCT-IV, observe that for $1 \le k \le \frac{m}{2}$, the jth entry of the column vector $\mathbf{c}_{-k}$ is

$$c(j, -k) = \cos \frac{(j + \frac{1}{2})(-k + \frac{1}{2})\pi}{n} = \cos \frac{(j + \frac{1}{2})(k - 1 + \frac{1}{2})\pi}{n} = c(j, k - 1),$$

so that $\mathbf{c}_{-k} = \mathbf{c}_{k-1}$, while for $k = n, \ldots, n + \frac{m}{2} - 1$, the jth entry of $\mathbf{c}_k$ is

$$c(j, k) = \cos \frac{(j + \frac{1}{2})(k + \frac{1}{2})\pi}{n} = -\cos \frac{(j + \frac{1}{2})(2n - k - \frac{1}{2})\pi}{n}$$
$$= -c(j, 2n - 1 - k),$$

so that $\mathbf{c}_{n+k} = -\mathbf{c}_{n-k-1}$ for $k = 0, \ldots, \frac{m}{2} - 1$. Hence, we have the following results on the computation of the $(m + n)$ square matrix BB^T:

(i) For $-\frac{m}{2} \le -k \le -1$, and $-\frac{m}{2} \le -j \le -1$,

$$\mathbf{c}_{-k}^T \mathbf{c}_{-j} = \mathbf{c}_{k-1}^T \mathbf{c}_{j-1} = \delta_{j-k}.$$

(ii) For $-\frac{m}{2} \le -k \le -1$, and $0 \le j \le \frac{m}{2} - 1$,

$$\mathbf{c}_{-k}^T \mathbf{c}_j = \mathbf{c}_{k-1}^T \mathbf{c}_j = \delta_{j-k+1}.$$

(iii) For $-\frac{m}{2} \le -k \le -1$, and $\frac{m}{2} \le j \le n + \frac{m}{2} - 1$,

$$\mathbf{c}_{-k}^T \mathbf{c}_j = 0.$$

(iv) For $n \le k \le n + \frac{m}{2} - 1$, and $n \le j \le n + \frac{m}{2} - 1$,

$$\mathbf{c}_k^T \mathbf{c}_j = \mathbf{c}_{2n-k-1}^T \mathbf{c}_{2n-j-1} = \delta_{j-k}.$$

(v) For $n \le k \le n + \frac{m}{2} - 1$, and $n - \frac{m}{2} \le j \le n - 1$,

$$\mathbf{c}_k^T \mathbf{c}_j = -\mathbf{c}_{2n-k-1}^T \mathbf{c}_j = -\delta_{2n-k-j-1}.$$

(vi) For $n \le k \le n + \frac{m}{2} - 1$ and $-\frac{m}{2} \le j \le n - \frac{m}{2} - 1$,

$$\mathbf{c}_k^T \mathbf{c}_j = 0.$$

Therefore, since the $(n + m)$-dimensional square matrix BB^T is given by $BB^T = [\mathbf{c}_k^T \mathbf{c}_j]$, $-\frac{m}{2} \le k, j \le n + \frac{m}{2} - 1$, and C_n is an orthogonal matrix,

$$BB^T = \begin{bmatrix} I_{\frac{m}{2}} & J_{\frac{m}{2}} & O \\ \cdots & \cdots & \cdots \\ O & I_n & O \\ \cdots & \cdots & \cdots \\ O & -J_{\frac{m}{2}} & I_{\frac{m}{2}} \end{bmatrix}$$

with identity diagonal, where $J_{\frac{m}{2}}$ denotes the matrix

$$\begin{bmatrix} & & 1 \\ & \cdot^{\cdot^{\cdot}} & \\ 1 & & \end{bmatrix}$$

with 1's on the northeast-southwest diagonal (anti-diagonal). Similarly, the same argument also yields

$$B_0 B_0^T = \begin{bmatrix} I_n & O & O \\ O & -J_{\frac{m}{2}} & I_{\frac{m}{2}} \end{bmatrix}, \quad B_1 B_1^T = \begin{bmatrix} I_{\frac{m}{2}} & J_{\frac{m}{2}} & O \\ O & O & I_n \end{bmatrix}$$

(see Exercise 5). Hence, the output (inverse DCT-IV) data in (5.3.12) are

$$\begin{cases} w_0^0 = u_0, \ldots, w_{n-\frac{m}{2}-1}^0 = u_{n-\frac{m}{2}-1}, \\ w_{n-\frac{m}{2}}^0 = \tilde{u}_{n-\frac{m}{2}}, \ldots, w_{n-1}^0 = \tilde{u}_{n-1}, \\ w_n^0 = \tilde{u}_n - \tilde{u}_{n-1}, w_{n+1}^0 = \tilde{u}_{n+1} - \tilde{u}_{n-2}, \ldots, \\ w_{n+\frac{m}{2}-1}^0 = \tilde{u}_{n+\frac{m}{2}-1} - \tilde{u}_{n-\frac{m}{2}}. \end{cases}$$

$$\begin{cases} \text{For } \ell = 1, \ldots, L - 2, \\ w_{\ell n-\frac{m}{2}}^\ell = \tilde{u}_{\ell n-\frac{m}{2}} + \tilde{u}_{\ell n+\frac{m}{2}-1}, \ldots, w_{\ell n-1}^\ell = \tilde{u}_{\ell n-1} + \tilde{u}_{\ell n}, \\ \\ w_{\ell n}^\ell = \tilde{u}_{\ell n}, \ldots, w_{(\ell+1)n-1}^\ell = \tilde{u}_{(\ell+1)n-1}, \\ \\ w_{(\ell+1)n}^\ell = \tilde{u}_{(\ell+1)n} - \tilde{u}_{(\ell+1)n-1}, \ldots, \\ \qquad w_{(\ell+1)n+\frac{m}{2}-1}^\ell = \tilde{u}_{(\ell+1)n+\frac{m}{2}-1} - \tilde{u}_{(\ell+1)n-\frac{m}{2}}, \end{cases}$$

$$\begin{cases} w^{L-1}_{(L-1)n-\frac{m}{2}} = \widetilde{u}_{(L-1)n-\frac{m}{2}} + \widetilde{u}_{(L-1)n+\frac{m}{2}-1}, \ldots, \\ \qquad w^{L-1}_{(L-1)n-1} = \widetilde{u}_{(L-1)n-1} + \widetilde{u}_{(L-1)n}, \\[2mm] w^{L-1}_{(L-1)n} = \widetilde{u}_{(L-1)n}, \ldots, w^{L-1}_{(L-1)n+\frac{m}{2}-1} = \widetilde{u}_{(L-1)n+\frac{m}{2}-1}, \\[2mm] w^{L-1}_{(L-1)n+\frac{m}{2}} = u_{(L-1)n+\frac{m}{2}}, \ldots, w^{L-1}_{Ln-1} = u_{Ln-1}. \end{cases} \qquad (5.3.13)$$

Finally, to recover the original data set $U = \{u_0, \ldots, u_{N-1}\} = \{u_0, \ldots, u_{Ln-1}\}$, we apply the result in (5.3.13) to the pre-processed data $\widetilde{u}_\ell = \{\widetilde{u}_{\ell n}, \ldots, \widetilde{u}_{(\ell+1)n-1}\}$, $\ell = 0, \ldots, L - 1$, as defined in (5.3.6)–(5.3.8), and simply invert 2×2 matrices, when necessary, as follows:

(i)

$$u_0 = w^0_0, \ldots, u_{n-\frac{m}{2}-1} = w^0_{n-\frac{m}{2}-1}$$

(ii) For $k = n - \frac{m}{2}, \ldots, n - 1$, apply the second result in the first set and first result in the second set of equations in (5.3.13) to yield

$$\begin{bmatrix} w^0_k \\ w^1_k \end{bmatrix} = \begin{bmatrix} h_k & -h_{2n-k-1} \\ (h_k + h_{k-n})\,(h_{n-k-1} - h_{2n-k-1}) \end{bmatrix} \begin{bmatrix} u_k \\ u_{2n-k-1} \end{bmatrix}$$

(iii) For $k = n, \ldots, n + \frac{m}{2} - 1$:

$$\begin{bmatrix} w^0_k \\ w^1_k \end{bmatrix} = \begin{bmatrix} (h_{k-n} + h_k)\,(h_{n-k-1} - h_{2n-k-1}) \\ h_{k-n} & h_{n-k-1} \end{bmatrix} \begin{bmatrix} u_k \\ u_{2n-k-1} \end{bmatrix}$$

(iv) For $\ell = 1, \ldots, L - 2;\ k = -\frac{m}{2}, \ldots, -1$:

$$w^\ell_{\ell n+k} = \begin{bmatrix} (h_{-k-1} - h_{n-k-1})\,(h_k + h_{n+k}) \end{bmatrix} \begin{bmatrix} u_{\ell n-k-1} \\ u_{\ell n+k} \end{bmatrix}$$

(v) For $\ell = 1, \ldots, L - 2;\ k = 0, \ldots, \frac{m}{2} - 1$:

$$w^\ell_{\ell n+k} = \begin{bmatrix} h_{-k-1} & h_k \end{bmatrix} \begin{bmatrix} u_{\ell n-k-1} \\ u_{\ell n+k} \end{bmatrix}$$

(vi) For $\ell = 1, \ldots, L - 2;\ k = \frac{m}{2}, \ldots, n - \frac{m}{2} - 1$:

$$w^\ell_{\ell n+k} = u_{\ell n+k}$$

(vii) For $\ell = 1, \ldots, L - 2; k = n - \frac{m}{2}, \ldots, n - 1$:

$$w^\ell_{\ell n+k} = \begin{bmatrix} h_k & -h_{2n-k-1} \end{bmatrix} \begin{bmatrix} u_{\ell n+k} \\ u_{\ell n+2n-k-1} \end{bmatrix}$$

(viii) For $\ell = 1, \ldots, L - 2; k = n, \ldots, n + \frac{m}{2} - 1$:

$$w^\ell_{\ell n+k} = \begin{bmatrix} (h_{k-n} + h_k) & (h_{n-k-1} - h_{2n-k-1}) \end{bmatrix} \begin{bmatrix} u_{\ell n+k} \\ u_{\ell n+2n-k-1} \end{bmatrix}$$

(ix) For $\ell = L - 1; k = -\frac{m}{2}, \ldots, -1$:

$$\begin{bmatrix} w^{\ell-1}_{\ell n+k} \\ w^\ell_{\ell n+k} \end{bmatrix} = \begin{bmatrix} -h_{n-k-1} & h_{n+k} \\ (h_{-k-1} - h_{n-k-1}) & (h_{n-k} + h_k) \end{bmatrix} \begin{bmatrix} u_{\ell n-k-1} \\ u_{\ell n+k} \end{bmatrix}$$

(x) For $\ell = L - 1; k = 0, \ldots, \frac{m}{2} - 1$:

$$\begin{bmatrix} w^{\ell-1}_{\ell n+k} \\ w^\ell_{\ell n+k} \end{bmatrix} = \begin{bmatrix} (h_{n-k-1} + h_{-k-1}) & (h_k - h_{n-k}) \\ h_{-k-1} & h_k \end{bmatrix} \begin{bmatrix} u_{\ell n-k-1} \\ u_{\ell n+k} \end{bmatrix}$$

(xi)

$$w^{L-1}_{(L-1)n+\frac{m}{2}} = u_{(L-1)n+\frac{m}{2}}, \ldots, w^{L-1}_{Ln-1} = u_{Ln-1}. \qquad \blacksquare$$

Remark 1 To understand why proper data pre-processing would help in suppressing blocky artifacts of quantized DCT, we may extend the notation of the two filter sequences H_1 and H_3 in (5.3.5) to formulate a "window sequence"

$$H = \{H_1, H_2, H_3\},$$

with $H_2 = \{1, \ldots, 1\}$ being the constant sequence of 1's and with length $= n - m$. Hence, the length of H is $n + m$. When the n-point DCT is applied to the pre-processed data segment, say $\widetilde{U}_\ell$ defined in (5.3.6)–(5.3.8), for DCT-IV, we have, for $\ell = 1, \ldots, L - 2$,

$$C_n \tilde{U}_\ell = \left[\mathbf{c}_0, \ldots, \mathbf{c}_{n-1} \right] \left[\tilde{u}_{\ell n}, \ldots, \tilde{u}_{(\ell+1)n-1} \right]^T$$

$$= \left[\mathbf{c}_{-\frac{m}{2}}, \ldots, \mathbf{c}_{n+\frac{m}{2}-1} \right] \left[h_{-\frac{m}{2}} u_{\ell n - \frac{m}{2}}, \ldots, h_{n+\frac{m}{2}-1} u_{(\ell+1)n+\frac{m}{2}-1} \right]^T,$$

since $\mathbf{c}_{-\frac{m}{2}} = \mathbf{c}_{\frac{m}{2}-1}, \ldots, \mathbf{c}_{-1} = \mathbf{c}_0$, $\mathbf{c}_n = -\mathbf{c}_{n-1}, \ldots, \mathbf{c}_{n+\frac{m}{2}-1} = -\mathbf{c}_{n-\frac{m}{2}}$.

Hence, by introducing the notion of "lapped orthogonal transform (LOT)"

$$A = \left[h_{-\frac{m}{2}} \mathbf{c}_{-\frac{m}{2}}, \ldots, h_{n+\frac{m}{2}-1} \mathbf{c}_{n+\frac{m}{2}-1} \right],$$

the n-point DCT of the pre-processed data is the same as the LOT of the original data, namely:

$$A \left[u_{\ell n - \frac{m}{2}}, \ldots, u_{(\ell+1)n+\frac{m}{2}-1} \right]^T = C_n \tilde{U}_\ell. \qquad \blacksquare$$

To better understand the lapped orthogonal transform, we next introduce the more general transformation without the orthogonality restriction. This will be called the "lapped transform". Let n be any positive integer, and $\{x_j\}$, $j \in \mathbb{Z}$, be a given bi-infinite data sequence. We partition this sequence into disjoint segments

$$X_\ell = \{x_{\ell n}, \ldots, x_{(\ell+1)n-1}\}, \qquad (5.3.14)$$

$\ell = 0, \pm 1, \ldots$, and extend each segment X_ℓ by tacking on m data values $x_{(\ell+1)n}, \ldots, x_{(\ell+1)n+m-1}$ (with $m \le n$ being any desirable positive integer) to formulate

$$X_\ell^{\text{ext}} = \{X_\ell^1, X_\ell^2, X_\ell^3\},$$

where

$$X_\ell^1 = \{x_{\ell n}, \ldots, x_{\ell n+m-1}\}, \qquad (5.3.15)$$
$$X_\ell^2 = \{x_{\ell n+m}, \ldots, x_{(\ell+1)n-1}\}, \qquad (5.3.16)$$
$$X_\ell^3 = \{x_{(\ell+1)n}, \ldots, x_{(\ell+1)n+m-1}\}. \qquad (5.3.17)$$

Observe that $X_\ell = \{X_\ell^1, X_\ell^2\}$ and X_ℓ^3 constitutes the extension X_ℓ^{ext} of X_ℓ.

Definition 1 *Let $0 < m \le n$ be integers and*

$$A = [A_1, A_2, A_3]$$

be an $n \times (n+m)$ matrix of real numbers, where A_1, A_2, A_3 are matrix sub-blocks of sizes $n \times m$, $n \times (n-m)$, $n \times m$, respectively. Suppose that there exists a corresponding $(n+m) \times n$ matrix

$$B = \begin{bmatrix} B_1 \\ B_2 \\ B_3 \end{bmatrix},$$

with matrix sub-blocks of B_1, B_2, B_3 of sizes $m \times n$, $(n-m) \times n$, $m \times n$, respectively, such that

$$
\begin{aligned}
B_j A_k &= O, \quad \text{for } j \neq k; \\
B_2 A_2 &= I_{n-m}, \\
B_1 A_1 + B_3 A_3 &= I_m,
\end{aligned}
\tag{5.3.18}
$$

where O denotes a zero matrix (with appropriate matrix dimensions depending on $j, k = 1, 2, 3$). Then A is called a lapped transform with dual transform B, and the pair (A, B) is called a (perfect reconstruction) dual pair.

When a lapped transform is applied to the extensions X_ℓ^{ext} of the disjoint segments X_ℓ of any sequence $\{x_j\}$, as discussed above, the dual transform can be used to recover X_ℓ for all $\ell \in \mathbb{Z}$, and hence the entire given sequence $\{x_j\}$. Indeed, since

$$A X_\ell^{\text{ext}} = A_1 X_\ell^1 + A_2 X_\ell^2 + A_3 X_\ell^3,$$

for such $\ell \in \mathbb{Z}$, we have

$$
\begin{bmatrix} Y_\ell^1 \\ Y_\ell^2 \\ Y_\ell^3 \end{bmatrix} = B(A X_\ell^{\text{ext}}) = \begin{bmatrix} B_1 A_1 X_\ell^1 + B_1 A_2 X_\ell^2 + B_1 A_3 X_\ell^3 \\ B_2 A_1 X_\ell^1 + B_2 A_2 X_\ell^2 + B_2 A_3 X_\ell^3 \\ B_3 A_1 X_\ell^1 + B_3 A_2 X_\ell^2 + B_3 A_3 X_\ell^3 \end{bmatrix}
$$

$$
= \begin{bmatrix} B_1 A_1 X_\ell^1 \\ B_2 A_2 X_\ell^2 \\ B_3 A_3 X_\ell^3 \end{bmatrix},
$$

or equivalently,

$$
\begin{aligned}
Y_\ell^1 &= B_1(A_1 X_\ell^1), \\
Y_\ell^2 &= I_{n-m} X_\ell^2 = X_\ell^2, \\
Y_\ell^3 &= B_3(A_3 X_\ell^3),
\end{aligned}
$$

for all $\ell \in \mathbb{Z}$. On the other hand, in view of the definition of the extension X_ℓ^{ext}, it is clear that

$$X_\ell^1 = X_{\ell-1}^3, \; \ell \in \mathbb{Z};$$

so that, in view of the third identity in (5.3.18),

$$Y_\ell^1 + Y_{\ell-1}^3 = B_1 A_1 X_\ell^1 + B_3 A_3 X_{\ell-1}^3 = (B_1 A_1 + B_3 A_3) X_\ell^1 = X_\ell^1.$$

Therefore, while X_ℓ^2 is identical to Y_ℓ^2, X_ℓ^1 can be recovered from $Y_\ell^1 + Y_{\ell-1}^3$, and hence each segment $X_\ell = \{X_\ell^1, X_\ell^2\}, \ell \in \mathbb{Z}$, of the sequence $\{x_j\}$ can be perfectly reconstructed. We summarize this result in the following.

Theorem 1 $\boxed{\textbf{Perfect reconstruction of lapped transform}}$ *Let $m \leq n$ be positive integers, $\{x_j\}$ be any bi-infinite sequence of real numbers, and $X_\ell = \{x_{\ell n}, \dots, x_{\ell n-1}\}$, $\ell \in \mathbb{Z}$. Also, let $X_\ell^{\text{ext}} = \{X_\ell, X_\ell^3\} = \{X_\ell^1, X_\ell^2, X_\ell^3\}$, where $X_\ell^1 = \{x_{\ell n}, \dots, x_{\ell n+m-1}\}$, $X_\ell^2 = \{x_{\ell n+m}, \dots, x_{(\ell+1)n-1}\}$, $X_\ell^3 = X_{\ell+1}^1 = \{x_{(\ell+1)n}, \dots, x_{(\ell+1)n+m-1}\}$. Suppose that A is a lapped transform with dual B, as introduced in Definition 1, then $\{x_j\}$ can be recovered from the transformed segments $A X_\ell^{\text{ext}}, \ell \in \mathbb{Z}$, by*

$$X_\ell^1 = B_1(A_1 X_\ell^1) + B_3(A_3 X_\ell^3)$$

and

$$X_\ell^2 = B_2(A_2 X_\ell^2),$$

where

$$X_\ell = \{X_\ell^1, X_\ell^2\}. \qquad\blacksquare$$

In Example 1, we have studied in some detail the recovery of the n-point DCT-IV of some pre-processed data, and in Remark 1, we have shown that such n-point DCT gives rise to the LOT of the original data. We will next study the more general windowed DCT by applying the dual lapped transform. Let $\mathbf{c}_k = [c(0, k), \dots, c(n-1, k)]^T$, $k = 0, \dots, n-1$, be the kth column of an $n \times n$ DCT matrix, such as DCT-I, DCT-II, DCT-III, or DCT-IV, introduced and studied in Sect. 4.4 of Chap. 4. (Recall that for DCT-I, C_n^I should be replaced by C_{n-1}^I so as to match the $n \times n$ dimension of the DCT matrix.) Now, extend the indices of the columns from $k = 0, \dots, n-1$ to $k = -\frac{m}{2}, \dots, n + \frac{m}{2} - 1$, where m is any desired positive even integer, with $2 \leq m \leq n$, and apply a window sequence

$$\left\{ h_{-\frac{m}{2}}, \dots, h_{n+\frac{m}{2}-1} \right\}$$

of real numbers with width (or length) $n + m$ to the extended DCT columns to formulate the $n \times (n+m)$ matrix

$$A = \left[h_{-\frac{m}{2}} \mathbf{c}_{-\frac{m}{2}}, \ldots, h_{n+\frac{m}{2}-1} \mathbf{c}_{n+\frac{m}{2}-1} \right] \tag{5.3.19}$$

(see Remark 1). Note that the kth column of A is

$$h_k \mathbf{c}_k = h_k [c(0, k), \ldots, c(n-1, k)]^T, \tag{5.3.20}$$

where $k = -\frac{m}{2}, \ldots, n + \frac{m}{2} - 1$. Before going ahead to restriction of the window sequence $\{h_{-\frac{m}{2}}, \ldots, h_{\frac{m}{2}+n-1}\}$ so that A, as defined in (5.3.19), is a lapped transform, let us first study the extension of the orthogonality property of the columns $\mathbf{c}_k$ of the DCT matrices, from $0 \le k \le n - 1$ to $-\frac{m}{2} \le k \le n + \frac{m}{2} - 1$, as follows. This will extend the study in Example 1 to include DCT-I, DCT-II, DCT-III.

Theorem 2 *Let m be an even integer and $2 \le m \le n$. Denote by*

$$C_n = [\mathbf{c}_0, \ldots, \mathbf{c}_{n-1}],$$

with columns $\mathbf{c}_0, \ldots, \mathbf{c}_{n-1}$, the DCT matrix for DCT-I, DCT-II, DCT-III, or DCT-IV (with n replaced by $n - 1$ for DCT-I, so that C_n is an $n \times n$ matrix). Then

$$\mathbf{c}_k^T \mathbf{c}_\ell = \langle \mathbf{c}_k, \mathbf{c}_\ell \rangle = \delta_{k-\ell} \tag{5.3.21}$$

for all $0 \le k, \ell \le n - 1$. Furthermore, the $n + m$ columns of the $n \times (n + m)$ matrix

$$C_n^{\text{ext}} = \left[\mathbf{c}_{-\frac{m}{2}} \ldots \mathbf{c}_0 \ldots \mathbf{c}_{n-1} \ldots \mathbf{c}_{n+\frac{m}{2}-1} \right] \tag{5.3.22}$$

satisfy the following properties:

(i) *For DCT-I, with $\mathbf{c}_k = \mathbf{c}_k^I$,*

$$\mathbf{c}_k = \mathbf{c}_\ell, \quad \text{for } k + \ell = 0 \text{ or } k + \ell = 2(n - 1),$$

where $-\frac{m}{2} \le k, \ell \le n + \frac{m}{2} - 1$.

(ii) *For DCT-II, with $\mathbf{c}_k = \mathbf{c}_k^{II}$,*

$$\mathbf{c}_k = \mathbf{c}_\ell, \quad \text{for } k + \ell = -1 \text{ or } k + \ell = 2n - 1,$$

where $-\frac{m}{2} \le k, \ell \le n + \frac{m}{2} - 1$.

(iii) *For DCT-III, with $\mathbf{c}_k = \mathbf{c}_k^{III}$,*

$$\mathbf{c}_k = \mathbf{c}_\ell, \quad \text{for } k + \ell = 0;$$
$$\mathbf{c}_k = -\mathbf{c}_\ell, \quad \text{for } k + \ell = 2n,$$

where $-\frac{m}{2} \le k, \ell \le n + \frac{m}{2} - 1$.

(iv) *For DCT-IV, with $\mathbf{c}_k = \mathbf{c}_k^{IV}$,*

$$\mathbf{c}_k = \mathbf{c}_\ell, \quad \text{for } k + \ell = -1;$$
$$\mathbf{c}_k = -\mathbf{c}_\ell, \quad \text{for } k + \ell = 2n - 1,$$

where $-\frac{m}{2} \leq k, \; \ell \leq n + \frac{m}{2} - 1.$

The orthogonality property (5.3.21) for the column vectors of DCT-I, ..., DCT-IV is guaranteed since such DCT matrices are orthogonal matrices (see Sect. 4.4 of Chap. 4). The symmetry properties at $x = 0$ for DCT-I and DCT-III and at $x = -\frac{1}{2}$ for DCT-II and DCT-IV are easy to verify. For the symmetry property at $x = n-1$ for DCT-I and at $x = n - \frac{1}{2}$ for DCT-II, see Exercise 6. For the anti-symmetry property at $x = n$ for DCT-III and at $x = n - \frac{1}{2}$ for DCT-IV (see Exercise 7 and Example 1). In view of these symmetry or anti-symmetry properties, the columns $\mathbf{c}_k$ of the matrix C_n^{ext} for $-\frac{m}{2} \leq k \leq -1$ and $n \leq k \leq n + \frac{m}{2} - 1$ inherit the orthogonality properties of the columns of C_n for $0 \leq \ell \leq \frac{m}{2} - 1$ and $n - \frac{m}{2} \leq \ell \leq n - 1$, respectively; of course except for $k + \ell = 0$ or $k + \ell = -1$, and for $k + \ell = 2n - 2$, or $k + \ell = 2n$, or $k + \ell = 2n - 1$. ∎

We now turn to the study of the $n \times (n + m)$ matrix A in (5.3.19), with column vectors

$$h_k \mathbf{c}_k, \quad k = -\frac{m}{2}, \ldots, n + \frac{m}{2} - 1,$$

given by (5.3.20), where

$$\mathbf{c}_k = [c(0, k), \ldots, c(n - 1, k)]^T,$$

and $c(j, k), \; j = 0, \ldots, n-1$, denotes the (j, k)-entry of the n-point DCT matrix C_n, which includes DCT-I, DCT-II, DCT-III, or DCT-IV. Our approach is to characterize the window sequence

$$\{h_{-\frac{m}{2}}, \ldots, h_{n+\frac{m}{2}-1}\}$$

in the definition of the $n \times (n + m)$ matrix A in (5.3.19) such that A is a lapped transform with dual $B = \widetilde{A}^T$, where

$$\widetilde{A} = \left[\widetilde{h}_{-\frac{m}{2}} \mathbf{c}_{-\frac{m}{2}}, \ldots, \widetilde{h}_{n+\frac{m}{2}-1} \mathbf{c}_{n+\frac{m}{2}-1}\right] = \left[\widetilde{A}_1, \widetilde{A}_2, \widetilde{A}_3\right] \tag{5.3.23}$$

for some sequence $\{\widetilde{h}_{-\frac{m}{2}}, \ldots, \widetilde{h}_{n+\frac{m}{2}-1}\}$, and $\widetilde{A}_1, \widetilde{A}_2, \widetilde{A}_3$ are $n \times m, n \times (n-m)$, and $n \times m$ sub-matrix blocks, respectively. For convenience, we call $\{\widetilde{h}_{-\frac{m}{2}}, \ldots, \widetilde{h}_{n+\frac{m}{2}-1}\}$ a dual window sequence. Observe that by setting

$$B = \begin{bmatrix} B_1 \\ B_2 \\ B_3 \end{bmatrix},$$

we have

$$B_1 = \left[\widetilde{h}_{-\frac{m}{2}} \mathbf{c}_{-\frac{m}{2}}, \ldots, \widetilde{h}_{\frac{m}{2}-1} \mathbf{c}_{\frac{m}{2}-1} \right]^T,$$

$$B_2 = \left[\widetilde{h}_{\frac{m}{2}} \mathbf{c}_{\frac{m}{2}}, \ldots, \widetilde{h}_{n-\frac{m}{2}-1} \mathbf{c}_{n-\frac{m}{2}-1} \right]^T, \qquad (5.3.24)$$

$$B_3 = \left[\widetilde{h}_{n-\frac{m}{2}} \mathbf{c}_{n-\frac{m}{2}}, \ldots, \widetilde{h}_{n+\frac{m}{2}-1} \mathbf{c}_{n+\frac{m}{2}-1} \right]^T,$$

and

$$B_1 A_1 = \widetilde{A}_1^T A_1, \ B_2 A_2 = \widetilde{A}_2^T A_2, \ B_3 A_3 = \widetilde{A}_3^T A_3.$$

From Definition 1, for the pair $(A, B) = (A, \widetilde{A}^T)$ to be a perfect reconstruction dual lapped transform pair, the window sequence $\{h_{-\frac{m}{2}}, \ldots, h_{n+\frac{m}{2}-1}\}$ and its dual $\{\widetilde{h}_{-\frac{m}{2}}, \ldots, \widetilde{h}_{n+\frac{m}{2}-1}\}$ must be so chosen that the three matrix equations in (5.3.18) are satisfied. That is, by setting

$$\widetilde{A}_\ell^T A_\ell = [a_{j,k}^\ell], \quad \ell = 1, 2, 3, \qquad (5.3.25)$$

the window and its corresponding dual window sequences must be selected to satisfy the following three conditions:

$$a_{j,k}^2 = \delta_{j-k}, \quad 1 \le j, \ k \le n - m; \qquad (5.3.26)$$

$$a_{j,k}^1 + a_{j,k}^3 = \delta_{j-k}, \quad 1 \le j, \ k \le m; \qquad (5.3.27)$$

$$\widetilde{A}_1^T A_2 = 0, \ \widetilde{A}_1^T A_3 = 0, \ \widetilde{A}_2^T A_3 = 0. \qquad (5.3.28)$$

Let us first observe that (5.3.28) always holds, independent of the choice of $\{h_{-\frac{m}{2}}, \ldots, h_{n+\frac{m}{2}-1}\}$ and $\{\widetilde{h}_{-\frac{m}{2}}, \ldots, \widetilde{h}_{n+\frac{m}{2}-1}\}$, since

$$\langle \widetilde{h}_k \mathbf{c}_k, h_\ell \mathbf{c}_\ell \rangle = \widetilde{h}_k h_\ell \langle \mathbf{c}_k, \mathbf{c}_\ell \rangle = 0$$

for $-\frac{m}{2} \le k \le \frac{m}{2} - 1$ and $\frac{m}{2} \le \ell \le n - \frac{m}{2} - 1$, or for $-\frac{m}{2} \le k \le \frac{m}{2} - 1$ and $n - \frac{m}{2} \le \ell \le n + \frac{m}{2} - 1$, or for $\frac{m}{2} \le k \le n - \frac{m}{2} - 1$ and $n - \frac{m}{2} \le \ell \le n + \frac{m}{2} - 1$, respectively, by applying the extension property in (iii)–(iv) in Theorem 2 and the orthogonality property (5.3.21).

To study the Eq. (5.3.26), we again apply (5.3.21) in Theorem 2 to conclude that

$$\widetilde{A}_2^T A_2 = \left[\langle \widetilde{h}_k \mathbf{c}_k, h_\ell \mathbf{c}_\ell \rangle \right]$$

$$= \left[\widetilde{h}_k h_\ell \langle \mathbf{c}_k, \mathbf{c}_\ell \rangle \right] = \widetilde{h}_k h_\ell \delta_{k-\ell}$$

for $\frac{m}{2} \le k, \ \ell \le n - \frac{m}{2} - 1$. Hence, (5.3.26) is satisfied, if and only if

$$\widetilde{h}_{\frac{m}{2}} h_{\frac{m}{2}} = \ldots = \widetilde{h}_{n-\frac{m}{2}-1} h_{n-\frac{m}{2}-1} = 1. \tag{5.3.29}$$

To study the Eq. (5.3.27), we compute $\widetilde{A}_1^T A_1$ and $\widetilde{A}_3^T A_3$ individually, as follows. For $1 \leq j, k \leq m$, we have

$$
\begin{aligned}
a_{j,k}^1 &= \langle \widetilde{h}_{j-\frac{m}{2}-1} \mathbf{c}_{j-\frac{m}{2}-1}, h_{k-\frac{m}{2}-1} \mathbf{c}_{k-\frac{m}{2}-1} \rangle \\
&= \widetilde{h}_{j-\frac{m}{2}-1} h_{k-\frac{m}{2}-1} \langle \mathbf{c}_{j-\frac{m}{2}-1}, \mathbf{c}_{k-\frac{m}{2}-1} \rangle,
\end{aligned}
\tag{5.3.30}
$$

and

$$
\begin{aligned}
a_{j,k}^3 &= \langle \widetilde{h}_{j+n-\frac{m}{2}-1} \mathbf{c}_{j+n-\frac{m}{2}-1}, h_{k+n-\frac{m}{2}-1} \mathbf{c}_{k+n-\frac{m}{2}-1} \rangle \\
&= \widetilde{h}_{j+n-\frac{m}{2}-1} h_{k+n-\frac{m}{2}-1} \langle \mathbf{c}_{j+n-\frac{m}{2}-1}, \mathbf{c}_{k+n-\frac{m}{2}-1} \rangle.
\end{aligned}
\tag{5.3.31}
$$

Observe that in view of (5.3.21) and properties in (i)–(iv) in Theorem 2, the inner products of the DCT column vectors in (5.3.30) and (5.3.31) are equal to 0, unless the two column vectors are the same, or unless one is an extended column in C_n^{ext} and the other is a column of C_n and they are either equal (due to symmetry) or the negative of each other (due to anti-symmetry). Hence, it is necessary to investigate each of the DCTs separately.

(i) For DCT-I, it follows from (i) in Theorem 2 that symmetry at $x = 0$ implies

$$\left(j - \frac{m}{2} - 1 \right) + \left(k - \frac{m}{2} - 1 \right) = 0,$$

or $j + k = m + 2$, and symmetry at $x = n - 1$ implies

$$\left(j + n - \frac{m}{2} - 1 \right) + \left(k + n - \frac{m}{2} - 1 \right) = 2(n - 1),$$

or $j + k = m$. Hence, in view of (5.3.21), we have

$$
a_{j,k}^1 + a_{j,k}^3 =
\begin{cases}
\widetilde{h}_{j-\frac{m}{2}-1} h_{j-\frac{m}{2}-1} + \widetilde{h}_{j+n-\frac{m}{2}-1} h_{j+n-\frac{m}{2}-1}, & \text{for } k = j, \\[2ex]
\widetilde{h}_{j-\frac{m}{2}-1} h_{\frac{m}{2}-j+1} + \widetilde{h}_{n+j-\frac{m}{2}-1} h_{\frac{m}{2}+n-j+1}, \\
\qquad\qquad\qquad\qquad \text{for } k = -j + m + 2, \\[2ex]
\widetilde{h}_{j-\frac{m}{2}-1} h_{\frac{m}{2}-j-1} + \widetilde{h}_{n+j-\frac{m}{2}-1} h_{\frac{m}{2}+n-j-1}, \\
\qquad\qquad\qquad\qquad \text{for } k = -j + m, \\[2ex]
0, & \text{otherwise.}
\end{cases}
$$

Consequently, for the window sequences $\{h_{-\frac{m}{2}}, \ldots, h_{n+\frac{m}{2}-1}\}$ and $\{\widetilde{h}_{-\frac{m}{2}}, \ldots,$
$\widetilde{h}_{n+\frac{m}{2}-1}\}$ to satisfy (5.3.27), it is necessary and sufficient that

$$\widetilde{h}_{j-\frac{m}{2}-1}h_{j-\frac{m}{2}-1} + \widetilde{h}_{j+n-\frac{m}{2}-1}h_{j+n-\frac{m}{2}-1} = 1, \text{ for } j = 2, \ldots, m-1;$$

$$\widetilde{h}_{j-\frac{m}{2}-1}h_{\frac{m}{2}-j+1} + \widetilde{h}_{j+n-\frac{m}{2}-1}h_{\frac{m}{2}+n-j+1} = 0, \text{ for } j = 1, \ldots, m;$$

$$\widetilde{h}_{j-\frac{m}{2}-1}h_{\frac{m}{2}-j-1} + \widetilde{h}_{j+n-\frac{m}{2}-1}h_{\frac{m}{2}+n-j-1} = 0, \text{ for } j = 1, \ldots, m.$$

(5.3.32)

(ii) For DCT-II, it follows from (ii) in Theorem 2 that symmetry at $x = -\frac{1}{2}$ implies

$$\left(j - \frac{m}{2} - 1\right) + \left(k - \frac{m}{2} - 1\right) = -1,$$

or $j + k = m + 1$, and that symmetry at $x = n - \frac{1}{2}$ implies

$$\left(j + n - \frac{m}{2} - 1\right) + \left(k + n - \frac{m}{2} - 1\right) = 2n - 1,$$

or $j + k = m + 1$ as well. Hence, in view of (5.3.21), we have

$$a_{j,k}^1 + a_{j,k}^3 = \begin{cases} \widetilde{h}_{j-\frac{m}{2}-1}h_{j-\frac{m}{2}-1} + \widetilde{h}_{j+n-\frac{m}{2}-1}h_{j+n-\frac{m}{2}-1}, & \text{for } k = j, \\[2ex] \widetilde{h}_{j-\frac{m}{2}-1}h_{\frac{m}{2}-j} + \widetilde{h}_{j+n-\frac{m}{2}-1}h_{\frac{m}{2}+n-j}, \\ \hspace{4cm} \text{for } k = -j+m+1, \\[2ex] 0, & \text{otherwise.} \end{cases}$$

Consequently, for the window sequences $\{h_{-\frac{m}{2}}, \ldots, h_{n+\frac{m}{2}-1}\}$ and $\{\widetilde{h}_{-\frac{m}{2}}, \ldots,$
$\widetilde{h}_{n+\frac{m}{2}-1}\}$ to satisfy (5.3.27), it is necessary and sufficient that

$$\widetilde{h}_{j-\frac{m}{2}-1}h_{j-\frac{m}{2}-1} + \widetilde{h}_{j+n-\frac{m}{2}-1}h_{j+n-\frac{m}{2}-1} = 1, \text{ for } j = 2, \ldots, m-1;$$

$$\widetilde{h}_{j-\frac{m}{2}-1}h_{\frac{m}{2}-j} + \widetilde{h}_{j+n-\frac{m}{2}-1}h_{\frac{m}{2}+n-j} = 0, \text{ for } j = 1, \ldots, m. \quad (5.3.33)$$

(iii) For DCT-III, it follows from (iii) in Theorem 2 that symmetry at $x = 0$ implies

$$\left(j - \frac{m}{2} - 1\right) + \left(k - \frac{m}{2} - 1\right) = 0,$$

or $j + k = m + 2$, and that anti-symmetry at $x = n$ implies

$$\left(j + n - \frac{m}{2} - 1\right) + \left(k + n - \frac{m}{2} - 1\right) = 2n,$$

or $j + k = m + 2$ as well. Hence, in view of (5.3.21), we have

$$a_{j,k}^1 + a_{j,k}^3 = \begin{cases} \widetilde{h}_{j-\frac{m}{2}-1}h_{j-\frac{m}{2}-1} + \widetilde{h}_{j+n-\frac{m}{2}-1}h_{j+n-\frac{m}{2}-1} \\ \qquad\qquad \times \left(1 - \delta_{j-\frac{m}{2}-1}\right), \quad \text{for } k = j, \\[2mm] \widetilde{h}_{j-\frac{m}{2}-1}h_{\frac{m}{2}-j+1} - \widetilde{h}_{j+n-\frac{m}{2}-1}h_{\frac{m}{2}+n-j+1}, \\ \qquad\qquad\qquad \text{for } k = -j + m + 2, \\[2mm] 0, \quad \text{otherwise.} \end{cases}$$

We remark that the factor $(1 - \delta_{j-\frac{m}{2}-1})$ of the term

$$\widetilde{h}_{j+n-\frac{m}{2}-1}h_{j+n-\frac{m}{2}-1}(1 - \delta_{j-\frac{m}{2}-1})$$

is needed, since the column $\mathbf{c}_n$ of C_n^{ext} for DCT-III is the zero-column (due to $\cos \frac{(j+\frac{1}{2})n\pi}{n} = 0, j = 0, \ldots, n - 1$). Consequently, for the sequences $\{h_{-\frac{m}{2}}, \ldots, h_{n+\frac{m}{2}-1}\}$ and $\{\widetilde{h}_{-\frac{m}{2}}, \ldots, \widetilde{h}_{n+\frac{m}{2}-1}\}$ to satisfy (5.3.27), it is necessary and sufficient that

$$\widetilde{h}_{j-\frac{m}{2}-1}h_{j-\frac{m}{2}-1} + \widetilde{h}_{j+n-\frac{m}{2}-1}h_{j+n-\frac{m}{2}-1} = 1, \text{ for } j \neq \frac{m}{2} + 1;$$

$$\widetilde{h}_0 h_0 = 1, \text{ for } j = \frac{m}{2} + 1;$$

$$\widetilde{h}_{j-\frac{m}{2}-1}h_{\frac{m}{2}-j+1} = \widetilde{h}_{j+n-\frac{m}{2}-1}h_{\frac{m}{2}+n-j+1}, \text{ for } j = 1, \ldots, m. \qquad (5.3.34)$$

(iv) For DCT-IV, it follows from (iv) in Theorem 2 that symmetry at $x = -\frac{1}{2}$ implies

$$\left(j - \frac{m}{2} - 1\right) + \left(k - \frac{m}{2} - 1\right) = -1,$$

or $j + k = m + 1$, and that anti-symmetry at $x = n - \frac{1}{2}$ implies

$$\left(j + n - \frac{m}{2} - 1\right) + \left(k + n - \frac{m}{2} - 1\right) = 2n - 1,$$

or $j + k = m + 1$ as well. Hence, in view of (5.3.21), we have

$$a^1_{j,k} + a^3_{j,k} = \begin{cases} \widetilde{h}_{j-\frac{m}{2}-1}h_{j-\frac{m}{2}-1} + \widetilde{h}_{j+n-\frac{m}{2}-1}h_{j+n-\frac{m}{2}-1}, & \text{for } k = j, \\[2ex] \widetilde{h}_{j-\frac{m}{2}-1}h_{\frac{m}{2}-j} - \widetilde{h}_{j+n-\frac{m}{2}-1}h_{\frac{m}{2}+n-j}, & \\[1ex] & \text{for } k = -j+m+1, \\[2ex] 0, & \text{otherwise.} \end{cases}$$

Consequently, for the sequences $\{h_{-\frac{m}{2}}, \ldots, h_{n+\frac{m}{2}-1}\}$ and $\{\widetilde{h}_{-\frac{m}{2}}, \ldots, \widetilde{h}_{n+\frac{m}{2}-1}\}$ to satisfy (5.3.27), it is necessary and sufficient that

$$\widetilde{h}_{j-\frac{m}{2}-1}h_{j-\frac{m}{2}-1} + \widetilde{h}_{j+n-\frac{m}{2}-1}h_{j+n-\frac{m}{2}-1} = 1, \text{ for } j = 2, \ldots, m-1;$$
$$\widetilde{h}_{j-\frac{m}{2}-1}h_{\frac{m}{2}-j} = \widetilde{h}_{j+n-\frac{m}{2}-1}h_{\frac{m}{2}+n-j}, \text{ for } j = 1, \ldots, m. \tag{5.3.35}$$

Remark 2 To apply the theory and methods of lapped transform to the data sequence $U = \{u_0, \ldots, u_{N-1}\}$, it is extended to a bi-infinite sequence by tacking on 0's both to the left and to the right of U. In addition, to adopt the notation of $AX^{\text{ext}} = A_1 X^1_\ell + A_2 X^2_\ell + A_3 X^3_\ell$ so that Theorem 1 can be applied, we set

$$\begin{aligned} x_k &= u_{k-\frac{m}{2}}, \text{ for } k = \tfrac{m}{2}, \ldots, N+\tfrac{m}{2}-1; \\ x_k &= 0, \qquad \text{for } k < \tfrac{m}{2} \text{ or } k \ge N+\tfrac{m}{2}. \end{aligned} \tag{5.3.36}$$

Also, by introducing the diagonal matrices

$$\begin{aligned} H_1 &= \text{diag } \{h_{-\frac{m}{2}}, \ldots, h_{\frac{m}{2}-1}\}, \\ \widetilde{H}_1 &= \text{diag } \{\widetilde{h}_{-\frac{m}{2}}, \ldots, \widetilde{h}_{\frac{m}{2}-1}\}, \\ H_2 &= \text{diag } \{h_{\frac{m}{2}}, \ldots, h_{n-\frac{m}{2}-1}\} = I_{n-m}, \\ \widetilde{H}_2 &= \text{diag } \{\widetilde{h}_{\frac{m}{2}}, \ldots, \widetilde{h}_{n-\frac{m}{2}-1}\} = H_2 = I_{n-m}, \\ H_3 &= \text{diag } \{h_{n-\frac{m}{2}}, \ldots, h_{n+\frac{m}{2}-1}\}, \\ \widetilde{H}_3 &= \text{diag } \{\widetilde{h}_{n-\frac{m}{2}}, \ldots, \widetilde{h}_{n+\frac{m}{2}-1}\}, \end{aligned}$$

we have

$$\begin{cases} A_1 = [\mathbf{c}_{-\frac{m}{2}}, \ldots, \mathbf{c}_{\frac{m}{2}-1}]H_1, \\ A_2 = [\mathbf{c}_{\frac{m}{2}}, \ldots, \mathbf{c}_{n-\frac{m}{2}-1}], \\ A_3 = [\mathbf{c}_{n-\frac{m}{2}}, \ldots, \mathbf{c}_{n+\frac{m}{2}-1}]H_3, \end{cases} \tag{5.3.37}$$

and

$$\begin{cases} B_1 = \widetilde{H}_1[\mathbf{c}_{-\frac{m}{2}}, \ldots, \mathbf{c}_{\frac{m}{2}-1}]^T, \\ B_2 = A_2^T, \\ B_3 = \widetilde{H}_3[\mathbf{c}_{n-\frac{m}{2}}, \ldots, \mathbf{c}_{n+\frac{m}{2}-1}]^T. \end{cases} \tag{5.3.38}$$

Hence, in view of (5.3.37), the diagonal matrices H_1, H_2, and H_3 are applied to the original data sequence segments for data pre-processing before the DCT is taken. Furthermore, in view of (5.3.38), the diagonal matrices $\widetilde{H}_1$, $\widetilde{H}_2$, and $\widetilde{H}_3$ are applied to recover the original data sequence segments after the inverse DCT is taken. Therefore the sequence $\{h_{-\frac{m}{2}}, \ldots, h_{n+\frac{m}{2}-1}\}$ is called a pre-processing filter, and the sequence $\{\widetilde{h}_{-\frac{m}{2}}, \ldots, \widetilde{h}_{n+\frac{m}{2}-1}\}$ (for the inverse DCT) is called a post-processing filter.   ∎

Example 2 For $n \geq m = 4$, let

$$h_{-2} = \frac{2-\sqrt{3}}{2\sqrt{2}}, \ h_{-1} = \frac{\sqrt{2}-1}{2}, \ h_0 = \frac{\sqrt{1+2\sqrt{2}}}{2}, \ h_1 = 1.$$

Verify that

$$H = \{h_{-2}, h_{-1}, h_0, h_1, 1, \ldots, 1, h_{n-2}, h_{n-1}, h_n, h_{n+1}\},$$

with $h_{n-2} = 1$, $h_{n-1} = h_0$, $h_n = h_{-1}$, $h_{n+1} = h_{-2}$, satisfies the condition (5.3.35) and hence, together with $\widetilde{H} = H$, guarantee that the window DCT A in (5.3.19) and window IDCT $B = [B_1, B_2, B_3]^T$ in (5.3.24) constitute a dual pair of lapped transform.

Solution In view of (5.3.35), since $\widetilde{h}_j = h_j$ for $j = -2, \ldots, n+1$, it is sufficient to verify that

$$h_{j-3}^2 + h_{j+n-3}^2 = 1, \ j = 2, 3,$$

and

$$h_{j-3}h_{2-j} = h_{j+n-3}h_{n-j+2}, \ j = 1, 2, 3, 4.$$

Indeed, by simple calculations, we have

$$\left(\frac{\sqrt{2}-1}{2}\right)^2 + \left(\frac{\sqrt{1+2\sqrt{2}}}{2}\right)^2 = \frac{2-2\sqrt{2}+1}{4} + \frac{1+2\sqrt{2}}{4} = 1$$

and the second set of conditions follows from the definition of $h_{n-2}, \ldots, h_{n+1}$. ∎

In our discussion of n-point DCT and the extension to lapped transforms, such as LOT for DCT-IV, we only considered 1-dimensional data sequences. To apply the DCT and lapped transform to data sets in higher dimensions, we may consider one dimension at a time. For example, to apply an n-point DCT to 2-dimensional data sets, such as digital images, we may first apply the transform in the horizontal direction, followed by the same transform in the vertical direction, as follows.

Definition 2 ┃2D-DCT┃ *Let A be an $m \times n$ data matrix. The **2-dimensional DCT** of A is defined by the n-point DCT C_n of the transpose A^T of A, followed by the*

m-point DCT C_m of the transpose of $C_n A^T$; so that the DCT of the data matrix A is defined by

$$\widehat{A} = C_m(C_n A^T)^T = C_m A C_n^T. \tag{5.3.39}$$

Furthermore, the corresponding inverse 2D-DCT is given by

$$A = C_m^T \widehat{A} C_n. \tag{5.3.40}$$

The reason for the need of taking matrix transposes in (5.3.39) is that in implementation, the 1-dimensional DCT is operated on rows of the data matrix.

Example 3 Compute the 2-dimensional DCT of the 2×2 matrix

$$A = \begin{bmatrix} 1 & 2 \\ -1 & 0 \end{bmatrix}.$$

Solution Recall from Example on p. 183 in Sect. 4.4 of Chap. 4 that the 2-point DCT is given by

$$C_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}.$$

Hence, it follows from (5.3.39) of Definition 2 that the 2-dimensional DCT of A is given by

$$
\begin{aligned}
\widehat{A} = C_2 A C_2^T &= \frac{1}{2} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \\
&= \frac{1}{2} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 3 & -1 \\ -1 & -1 \end{bmatrix} \\
&= \frac{1}{2} \begin{bmatrix} 2 & -2 \\ 4 & 0 \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ 2 & 0 \end{bmatrix}. \qquad \blacksquare
\end{aligned}
$$

Example 4 Compute the 2-dimensional DCT of the rectangular matrix

$$A = \begin{bmatrix} -1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}.$$

Solution Recall from Example 1 on p. 183 again that the 3-point DCT is given by

$$C_3 = \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & 0 & \frac{-1}{\sqrt{2}} \\ \frac{1}{\sqrt{6}} & -\sqrt{\frac{2}{3}} & \frac{1}{\sqrt{6}} \end{bmatrix}.$$

Hence, it follows from (5.3.39) of Definition 2 that the 2-dimensional DCT of A is given by

$$\widehat{A} = C_2 A C_3^T = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} -1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} C_3^T$$

$$= \frac{1}{\sqrt{2}} \begin{bmatrix} 0 & 1 & 1 \\ -2 & -1 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & 0 & -\sqrt{\frac{2}{3}} \\ \frac{1}{\sqrt{3}} & \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \end{bmatrix}$$

$$= \frac{1}{\sqrt{2}} \begin{bmatrix} \frac{2}{\sqrt{3}} & \frac{-1}{\sqrt{2}} & \frac{-1}{\sqrt{6}} \\ \frac{-2}{\sqrt{3}} & \frac{-3}{\sqrt{2}} & \frac{1}{\sqrt{6}} \end{bmatrix} = \begin{bmatrix} \sqrt{\frac{2}{3}} & \frac{-1}{2} & \frac{-1}{2\sqrt{3}} \\ -\sqrt{\frac{2}{3}} & \frac{-3}{2} & \frac{1}{2\sqrt{3}} \end{bmatrix}. \qquad \blacksquare$$

To formulate the 2-dimensional DCT and inverse DCT in terms of sums of products (without matrix-to-matrix multiplications), we recall from (4.2.12)–(4.2.13) on p.186 that the jth row of the DCT is given by

$$\mathbf{c}_j^T = d_j \sqrt{\frac{2}{n}} \left[\cos \frac{j\pi}{2n} \quad \cos \frac{j3\pi}{2n} \dots \cos \frac{j(2n-1)\pi}{2n} \right]$$

for $j = 0, \dots, n-1$, where

$$d_0 = \frac{1}{\sqrt{2}}; \quad d_1 = \dots = d_{n-1} = 1.$$

Therefore, it follows from the definition of the DCT $\widehat{A}$ of an $n \times n$ square matrix A and inverse DCT A of $\widehat{A}$, in (5.3.39)–(5.3.40) respectively, where

$$A = [a_{j,k}]_{0 \le j,k \le n-1}; \quad \widehat{A} = [\widehat{a}_{\ell,s}]_{0 \le \ell,s \le n-1},$$

that

$$\widehat{a}_{j,k} = \frac{2}{n} \sum_{\ell=0}^{n-1} \sum_{s=0}^{n-1} \left(d_j \cos \frac{j(2\ell-1)\pi}{2n} \right) a_{\ell,s} \left(d_k \cos \frac{k(2s-1)\pi}{2n} \right)$$

$$= \frac{2}{n} d_j d_k \sum_{\ell=0}^{n-1} \sum_{s=0}^{n-1} \left(\cos \frac{j(2\ell-1)\pi}{2n} \quad \cos \frac{k(2s-1)\pi}{2n} \right) a_{\ell,s}, \qquad (5.3.41)$$

for $j, k = 0, 1, \dots, n-1$; and

$$a_{\ell,s} = \frac{2}{n} \sum_{j=0}^{n-1} \sum_{k=0}^{n-1} \left(d_j \cos \frac{j(2\ell-1)\pi}{2n} \right) \widehat{a}_{j,k} \left(d_k \cos \frac{k(2s-1)\pi}{2n} \right)$$

$$= \frac{2}{n} \sum_{j=0}^{n-1} \sum_{k=0}^{n-1} \left(d_j d_k \cos \frac{j(2\ell-1)\pi}{2n} \quad \cos \frac{k(2s-1)\pi}{2n} \right) \widehat{a}_{j,k}, \qquad (5.3.42)$$

for $\ell, s = 0, \dots, n-1$.

Exercises

Exercise 1 Compute the quantization of

$$x_1 = 23 \text{ and } x_2 = -23$$

by applying (5.3.1) with quantizers $Q_1 = 2$, $Q_2 = 5$. Then compute the errors

$$x_1 - Q_1\tilde{x}_1, \ x_1 - Q_2\tilde{x}_1, \ x_2 - Q_1\tilde{x}_2, \ x_2 - Q_2\tilde{x}_2.$$

Exercise 2 Repeat Exercise 1 but use the quantization formula (5.3.2) instead.

Exercise 3 Compare the performance of the quantization formulas (5.3.1) and (5.3.2) by using more samples x than x_1 and x_2 and more quantizers Q than Q_1 and Q_2 in Exercise 1.

Exercise 4 To better understand Example 1, work at the example for $n = 4$ and $m = 2$. Verify that the original data can be recovered via the inverse DCT-IV C_4^T for the pre-processing scheme in Example 1.

Exercise 5 Show that the matrix B_0 and B_1 defined in (5.3.11) have the properties

$$B_0 B_0^T = \begin{bmatrix} I_n & O & O \\ O & -J_{\frac{m}{2}} & I_{\frac{m}{2}} \end{bmatrix}, \quad B_1 B_1^T = \begin{bmatrix} I_{\frac{m}{2}} & J_{\frac{m}{2}} & O \\ O & O & I_n \end{bmatrix},$$

where $J_{\frac{m}{2}}$ denotes the matrix

$$\begin{bmatrix} & & 1 \\ & \cdot^{\cdot^{\cdot}} & \\ 1 & & \end{bmatrix}.$$

Exercise 6 Verify the symmetry property at $x = n - 1$ for DCT-I and at $x = n - \frac{1}{2}$ for DCT-II.

Exercise 7 Verify the anti-symmetry property at $x = n - 1$ for DCT-III and at $x = n - \frac{1}{2}$ for DCT-IV.

Exercise 8 Compute the 2-dimensional DCT of the following matrices.

(a)

$$A_1 = \begin{bmatrix} 2 & 3 \\ 1 & 1 \end{bmatrix}.$$

(b)

$$A_2 = \begin{bmatrix} 2 & 0 & 3 \\ 1 & -1 & 0 \end{bmatrix}.$$

(c)

$$A_3 = \begin{bmatrix} 1 & 0 \\ 0 & 3 \\ 1 & 1 \end{bmatrix}.$$

(d)

$$A_4 = \begin{bmatrix} -1 & 0 & 1 \\ 1 & 1 & -1 \\ 0 & 0 & 1 \end{bmatrix}.$$

Exercise 9 Verify the solutions of Exercise 8 by computing the inverse DCT (or IDCT) of $\widehat{A}_1, \widehat{A}_2, \widehat{A}_3, \widehat{A}_4$, where $\widehat{A}_j$ is the DCT of A_j for $j = 1, \ldots, 4$ in Exercise 8.

Exercise 10 Apply the following quantizers q_j to the corresponding data to obtain the closest integers $(a_{j,k}/q_j)$.

(a) Quantizer: $q_1 = 4$;
    Data: $a_{1,0} = 161, a_{1,1} = 125, a_{1,2} = -30$.
(b) Quantizer: $q_2 = 3$;
    Data: $a_{2,0} = -162, a_{2,1} = 112, a_{2,2} = 17$.
(c) Quantizer: $q_3 = 60$;
    Data: $a_{3,0} = 150, a_{3,1} = -251, a_{3,2} = -180$.

Exercise 11 De-quantize each of the quantized integers $(a_{j,k}/q_j)$, $j = 1, 2, 3$, for the data $a_{j,k}$, $j = 1, 2, 3$ and $k = 0, 1, 2$ in (a), (b), (c) of Exercise 10, by computing

$$\widetilde{b}_{j,k} = q_j(a_{j,k}/q_j), \quad j = 1, 2, 3, k = 0, 1, 2.$$

Then compute the relative errors

$$\frac{a_{j,k} - \widetilde{b}_{j,k}}{a_{j,k}} \times 100\%,$$

for $j = 1, 2, 3$, and $k = 0, 1, 2$.

5.4 Image and Video Compression

In this section, we discuss the industry standards for image and video compressions. For image compression, the standard is called JPEG, which is the acronym for "Joint Photographic Experts Group". The compression scheme has been described in the previous section, with the transformation being the DCT applied to 8×8 tiles of the digital image. In other words, we apply (5.3.41)–(5.3.42) to 8×8 data matrices, as in the following result.

Theorem 1 *For each* 8×8 *sub-block*

$$
A = \begin{bmatrix} a_{0,0} & \cdots & a_{0,7} \\ \vdots & & \\ & \cdots & \\ a_{7,0} & \cdots & a_{7,7} \end{bmatrix}
\tag{5.4.1}
$$

of a digital image, the DCT

$$
\widehat{A} = \begin{bmatrix} \widehat{a}_{0,0} & \cdots & \widehat{a}_{0,7} \\ \vdots & & \\ & \cdots & \\ \widehat{a}_{7,0} & \cdots & \widehat{a}_{7,7} \end{bmatrix}
$$

of A is given by

$$
\widehat{a}_{j,k} = \frac{d_j d_k}{4} \sum_{\ell=0}^{7} \sum_{s=0}^{7} \left(\cos \frac{j(2\ell-1)\pi}{16} \cos \frac{k(2s-1)\pi}{16} \right) a_{\ell,s},
\tag{5.4.2}
$$

for $j, k = 0, \ldots, 7$; *and the IDCT of* $\widehat{A}$ *is given by*

$$
a_{\ell,s} = \frac{1}{4} \sum_{j=0}^{7} \sum_{k=0}^{7} \left(d_j d_k \cos \frac{j(2\ell-1)\pi}{16} \cos \frac{k(2s-1)\pi}{16} \right) \widehat{a}_{j,k},
\tag{5.4.3}
$$

for $\ell, s = 0, \ldots, 7$. *In* (5.4.2) *and* (5.4.3), $d_0 = \frac{1}{\sqrt{2}}$ *and* $d_1 = \cdots = d_7 = 1$.

Remark 1 For 8-bit images, the entries $a_{\ell,s}$ of the 8×8 image block A in (5.4.1) are integers that range from 0 to 255 (that is, $0 \leq a_{\ell,s} \leq 255$). Hence, it follows from (5.4.2) that the dc (direct current) term $\widehat{a}_{0,0}$ of the DCT of A is given by

$$
\widehat{a}_{0,0} = \frac{1}{8} \sum_{\ell=0}^{7} \sum_{s=0}^{7} a_{\ell,s},
$$

which is some integer between 0 and 2040 and therefore is a 11-bit integer. In addition, all the other DCT coefficients $\widehat{a}_{j,k}$ may oscillate in signs and are called ac (alternate current) terms. ∎

Remark 2 To reduce the size of the dc term $\widehat{a}_{0,0}$, DPCM is used in JPEG by taking the difference with the dc term of the previous 8×8 block. On the other hand, to achieve much higher compression ratio, the ac terms $\widehat{a}_{j,k}$ (for $(j, k) \neq (0, 0)$) are to be "quantized". This is the lossy (or non-reversible) encoding component of the JPEG encoder. The quantization process is to divide each $\widehat{a}_{j,k}$ by some positive
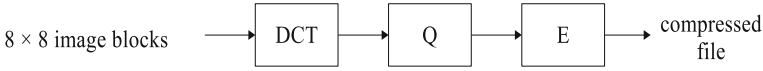
Fig. 5.1 Encoder : Q = quantization; E = entropy encoding
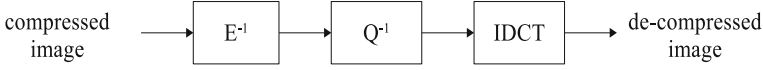
Fig. 5.2 Decoder : Q^{-1} = de-quantization; E^{-1} = de-coding

integer $q_{j,k}$, followed by rounding off the quotient to the nearest integer, to be denoted by $(\widehat{a}_{j,k}/q_{j,k})$. For instance, for large values of $q_{j,k}$, the round-off integers $(\widehat{a}_{j,k}/q_{j,k})$ are equal to 0, yielding a long row of consecutive zeros (0's) for high-frequency ac terms (that is, relatively larger values of $j + k$). The division by positive integers $q_{j,k}$ also yield smaller numbers (and hence, less bits) to encode. When decoding the (encoded) compressed image file, the integers $q_{j,k}$ are multiplied to the round-off integers $(\widehat{a}_{j,k}/q_{j,k})$ before the inverse DCT (denoted by IDCT) is applied. ∎

The schematic diagrams of the JPEG encoder and decoder are shown in Figs. 5.1 and 5.2.

In Figs. 5.3–5.4, we give two examples of the 8×8 quantizers, with one for achieving low compression ratio and one for achieving high compression ratio.

The round-off integers, denoted by

$$b_{j,k} = (a_{j,k}/q_{j,k}),$$

for $j, k = 0, \ldots, 7$, are arranged as a sequence of 64 integers by following the zig-zag ordering as shown in Fig. 5.5 and written out precisely as in (5.4.4).

1	1	1	1	1	2	2	4
1	1	1	1	1	2	2	4
1	1	1	1	2	2	2	4
1	1	1	1	1	2	4	8
1	1	2	2	2	2	4	8
2	2	2	2	2	4	8	8
2	2	2	4	4	8	8	16
4	4	4	4	8	8	16	16

Fig. 5.3 Quantizers: low compression ratio

1	2	4	8	16	32	64	128
2	4	4	8	16	32	64	128
4	4	8	16	32	64	128	128
8	8	16	32	64	128	128	256
16	16	32	64	128	128	256	256
32	32	64	128	128	256	256	256
64	64	128	128	256	256	256	256
128	128	128	256	256	256	256	256

Fig. 5.4   Quantizers: high compression ratio

Fig. 5.5   Zig-zag ordering

$$
\begin{aligned}
B = \{ & b_{0,0}, b_{0,1}, b_{1,0}, b_{2,0}, b_{1,1}, b_{0,2}, b_{0,3}, b_{1,2}, b_{2,1}, b_{3,0}, b_{4,0}, \\
& b_{3,1}, b_{2,2}, b_{1,3}, b_{0,4}, b_{0,5}, b_{1,4}, b_{2,3}, b_{3,2}, b_{4,1}, b_{5,0}, b_{6,0}, \\
& b_{5,1}, b_{4,2}, b_{3,3}, b_{2,4}, b_{1,5}, b_{0,6}, b_{0,7}, b_{1,6}, b_{2,5}, b_{3,4}, b_{4,3}, \\
& b_{5,2}, b_{6,1}, b_{7,0}, b_{7,1}, b_{6,2}, b_{5,3}, b_{4,4}, b_{3,5}, b_{2,6}, b_{1,7}, b_{2,7}, \\
& b_{3,6}, b_{4,5}, b_{5,4}, b_{6,3}, b_{7,2}, b_{7,3}, b_{6,4}, b_{5,5}, b_{4,6}, b_{3,7}, b_{4,7}, \\
& b_{5,6}, b_{6,5}, b_{7,4}, b_{7,5}, b_{6,6}, b_{5,7}, b_{6,7}, b_{7,6}, b_{7,7} \}.
\end{aligned}
\tag{5.4.4}
$$

Then B is considered as an information source to be encoded. The following modification is specified when the Huffman table, provided by the JPEG standard, is applied.

(i) The dc term $b_{0,0}$ in (5.4.4) is subtracted from the dc term of the previous 8×8 (DCT-quantized) block, and the result $\tilde{b}_{0,0}$ replaces $b_{0,0}$ of (5.4.4) as the modified information source. (Recall that this differencing step is called DPCM.)

(ii) Two code-words from the Huffman table are reserved for encoding rows of 0's in the sequence B in (5.4.4):

    (a) End-of-block (EOB) with code-word 00000000 (of 8 zeros) means that the remaining source words in B are zeros and there is no need to encode them. For example, in (5.4.4), if $b_{7,1} \neq 0$ but all the 27 source words in the sequence B after $b_{7,1}$ are equal to 0, then after encoding $b_{7,1}$ by applying the Huffman table, tack eight 0's to indicate $b_{7,1}$ is the last word to be encoded in this 8×8 block;

    (b) Zero run length (ZRL) with code-word 11110000 (of four 1's followed by four 0's) means that there are words of consecutive zeros in B before the last non-zero source word. We will not go into details of this process.

In the above discussion, we only considered compression of 8-bit gray-scale images, for convenience. For 24-bit RGB color images, with 8-bit R (red), 8-bit G (green), and 8-bit B (blue), it is recommended to apply the color transform from RGB to YC_bC_r, where Y stands for luminance (that is, light intensity, which by itself can be used to display the gray-scale image). The other two color components C_b and C_r, called chrominance (more precisely, chrominance blue and chrominance red, respectively), convey the color information by carrying the difference in intensities from the intensity of the luminance.

During the composite (analog) TV era, the YIQ color coordinates were introduced by RCA in the 1950s for broadcast bandwidth saving by "chroma-subsampling" of the IQ color components. It was later adopted by the NTSC standard, with the luminance Y specified to be

$$Y = 0.299R + 0.587G + 0.114B, \tag{5.4.5}$$

and chrominance I and Q to be

$$\begin{aligned} I &= 0.736(R - Y) - 0.268(B - Y), \\ Q &= 0.478(R - Y) + 0.413(B - Y). \end{aligned} \tag{5.4.6}$$

To overcome certain shortcomings, particularly in up-sampling, Germany introduced the PAL standard in the 1960s with YUV color transform given by

$$Y = 0.3R + 0.6G + 0.1B, \tag{5.4.7}$$

and the chrominance components U and V given simply by

$$U = B - Y, \ V = R - Y. \tag{5.4.8}$$

Meanwhile, France also introduced another standard called SECAM. The importance of the luminance–chrominance formats (as opposed to the RGB color coordinates) is that human vision is much more sensitive to light intensity (or brightness) than color differences, particularly for scenes in motion. Observe that both I, Q in (5.4.6) and particularly U, V in (5.4.8) are defined by taking color differences. Consequently, chroma-subsampling by down-sampling the chrominance components is hardly noticeable in digital TV broadcasting. Currently, the so-called 4:1:1 and 4:2:0 formats allow an additional 33 % increase in compression ratio. Furthermore, if noise removal is applied appropriately to the chrominance components, the Y component maintains the sharpness of the video imagery.

To adopt the YIQ and YUV color coordinates for color image compression, the YC_bC_r format was developed by the JPEG image compression standard, by specifying

$$\begin{array}{l} C_b = \frac{1}{2}\, U + 0.5; \\ C_r = \frac{1}{1.6}\, V + 0.5, \end{array} \tag{5.4.9}$$

for the chrominance components, where the values of the colors R, G, B are expressed by a relative scale from 0 to 1, with 0 indicating no phosphor excitation and 1 indicating maximum phosphor excitation. Hence, the additive factor of 0.5, with the decrease in the color differences $U = B - Y$ and $V = R - Y$, facilitates a better preservation of the blue and red colors, even after down-sampling of C_b and C_r to achieve higher compression ratio.

We end this chapter by giving a very brief introduction to (digital) video compression.

While the image compression standard JPEG was developed by the "Joint Photographic Experts Group" under the auspices of the three major international standard organizations ISO, CCITT, and IEC, the video compression standard MPEG was developed by the "Moving Pictures Expert Group" of ISO. The first standard completed in 1991 is known as MPEG-1 for the first-generation digital video compression. The second-generation digital video standard, known as MPEG-2, is currently adopted for HDTV broadcasting. In addition, MPEG-4 was developed in the late 1990's for low bit-rate video compression. More recently, the new video standard H.264, also called MPEG-4 Part 10 and AVC (for Advanced Video Coding), was successfully developed in 2003 for up to an additional 50 % compression saving over MPEG-2 and MPEG-4 Part 1, while maintaining comparable video quality. H.264 is the video compression standard for Blu-ray discs, and is widely used for internet video streaming by such giants as YouTube (of Google) and iTune stores (of Apple). In addition, it is embedded in the web software Adobe Flash Player, and is the preferred video format for most cable and satellite television service providers, including Direct TV.

In any case, all effective video compression schemes are similar, with I-pictures (or I-frames), P-pictures (or P-frames), and B-pictures (or B-frames). With the

exception of H.264, all MPEG and other H.26x standards adopt the JPEG image compression standard for I-picture (or intra-frame) compression. To meet the mandate of $2\times$ compression efficiency over MPEG-2, I-frame image compression for H.264 departs from the JPEG standard by introducing "I-slices" that include 4×4 DCT blocks, with applications to frame-by-frame video such as "iFrame video", developed by Apple in 2009 to facilitate video editing and high-quality video camcorder recording, particularly in iMovie'09 and iMovie'11. It is also adopted by camera manufacturers to capture HD video in 1920×1080 resolution, such as the AVC-Intra video codec (that is, encoding and decoding), developed by Panasonic in 2007 for HD video broadcasting.

On the other hand, to facilitate video frame prediction (for P-frames and B-frames), the I-frame format for video encoding is slightly different from JPEG in that the I-frames are partitioned into macroblocks of sizes 8×8 or 16×16. In other words, DPCM encoding of the dc coefficients is limited to at most four 8×8 blocks. To encode a P-frame (or prediction frame), adjacent macroblocks of the current P-frame (called intra-macroblocks) are compared with macroblocks of previous I or P frames (called inter-macroblocks) by "motion search". If an inter-macroblock (from some previous frame) is suitable to replace an adjacent intra-macroblock, then compression of this adjacent macroblock is eliminated simply by coding the "motion vector" that tells the decoder which inter-macroblock is used to replace the adjacent macroblock of the current frame. Bi-directional frame prediction is an extension of the P-frame prediction to allow searching of inter-macroblock replacement (for replacing adjacent macroblocks of the current frame) from both previous video frames and future video frames. Such prediction frames are called B-frames (or bi-directional prediction frames).

Exercises

Exercise 1 Show that the YIQ color coordinates can be computed approximately from the YUV color coordinates by a rotation of $33°$; that is,

$$\begin{bmatrix} I \\ Q \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \cos 33° & \sin 33° \\ -\sin 33° & \cos 33° \end{bmatrix} \begin{bmatrix} U \\ V \end{bmatrix}.$$

Exercise 2 Recover the original color coordinate (R, G, B) from (Y, U, V) by applying (5.4.7) and (5.4.8).

Exercise 3 Recover the original color coordinate (R, G, B) from (Y, I, Q) by applying (5.4.5) and (5.4.6).

Exercise 4 Recover the original color coordinate (R, G, B) from (Y, C_b, C_r) by applying (5.4.9).

Exercise 5 For a 24-bit color image with integers r, g, b, where $0 \leq r, g, b \leq 255$, let

$$R = \frac{r}{255}, \quad G = \frac{g}{255}, \quad B = \frac{b}{255}.$$

Apply (5.4.9) to compute the 24-bit YC_bC_r luminance /chrominance representation of (r, g, b).

Exercise 6 Repeat Exercise 5 by computing the YIQ luminance /chrominance representation of (r, g, b).

Exercise 7 Repeat Exercise 6 by computing the YUV luminance /chrominance representation of (r, g, b).

Chapter 6
Fourier Series
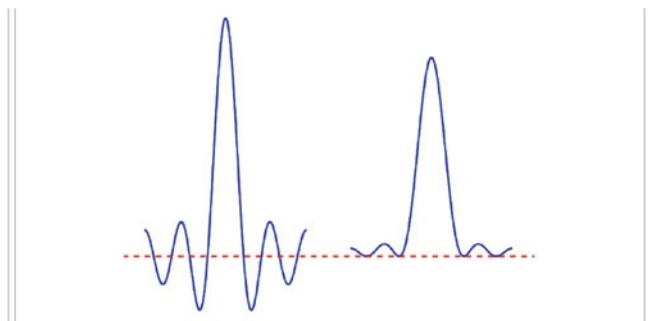
When an n-dimensional vector $\mathbf{x} = (x_0, \ldots, x_{n-1})$ in $\mathbb{C}^n$ represents a digital signal, the ordered set $\{0, \ldots, n-1\}$ of indices is called the discrete time-domain of the signal. Similarly, when a piecewise continuous function $f(x)$ defined on some bounded interval $[a, b]$ represents an analog signal, then the interval $[a, b]$ is called the continuous time-domain of $f(x)$. Such functions can be extended to periodic functions on the time-domain $\mathbb{R} = (-\infty, \infty)$ with period $= (b - a)$, by setting

$$f\big(x + \ell(b - a)\big) = f(x), \quad x \in [a, b],$$

for all $\ell = \pm 1, \pm 2, \ldots$, after replacing the values $f(a)$ and $f(b)$ by their average value $\big(f(a) + f(b)\big)/2$.

In the first section of this chapter, the analogy of discrete Fourier transform (DFT) of $\mathbf{x} \in \mathbb{C}^n$ studied in Chap. 4 is replaced by the definition of "Fourier-coefficients"

$$c_k(f) = \frac{1}{b - a} \int_a^b f(x) e^{-i2\pi k(x-a)/(b-a)} \, dx, \quad k \in \mathbb{Z},$$

of such periodic functions. Observe that while the n-point DFT $\widehat{\mathbf{x}} = \mathbb{F}_n \mathbf{x}$ of $\mathbf{x} \in \mathbb{C}^n$ is again a vector in $\mathbb{C}^n$, the sequence $\{c_k(f)\}$ of Fourier coefficients of $f \in PC[a, b]$ is an infinite (more precisely, a bi-infinite) sequence. Furthermore, in the view of Euler's formula,

$$e^{-i2\pi k(x-a)/(b-a)} = \cos \frac{2\pi k(x - a)}{b - a} - i \sin \frac{2\pi k(x - a)}{b - a},$$

the sequence $\{c_k(f)\}$ of Fourier coefficients of $f \in PC[a, b]$ reveals the frequency content of the analog signal $f(x)$, in a similar manner as the DFT $\widehat{\mathbf{x}}$ of $\mathbf{x}$ reveals the frequency content of a digital signal.

To recover $f(x)$ from its Fourier coefficients, the notion of Fourier series

$$(Sf)(x) = \sum_{k=-\infty}^{\infty} c_k(f)e^{i2\pi k(x-a)/(b-a)}$$

is also introduced in Sect. 4.1, with its partial sums defined by

$$(S_n f)(x) = \sum_{k=-n}^{n} c_k(f)e^{i2\pi k(x-a)/(b-a)}$$

for $n = 0, 1, \ldots$. As a first step toward the recovery of $f(x)$ from its Fourier series $(Sf)(x)$, it is shown in this section that the partial sums $(S_n f)(x)$ provide the best $L_2[a, b]$ approximation of $f(x)$.

For practical applications, it is often more convenient to get rid of the imaginary unit "i" in the Fourier series representation by re-formulating $(Sf)(x)$ in terms of cosine and/or sine. This is the topic of discussion in Sect. 6.2, where the same Fourier series $(Sf)(x)$ is formulated as

$$(Sf)(x) = \frac{a_0}{2} + \sum_{k=1}^{\infty} \left(a_k \cos \frac{2\pi k(x - a)}{b - a} + b_k \sin \frac{2\pi k(x - a)}{b - a} \right).$$

Pure cosine and pure sine series are also derived by even and odd extensions, respectively. In addition, the continuous version called "Type II cosine series" of the DCT, studied in Sect. 4.3 of Chap. 4 and applied to image and video compression in Chap. 5, is derived in this second section.

For convenience, we will only consider, without loss of generality, the interval $[a, b] = [-\pi, \pi]$ in Sect. 6.1 as well as Sects. 6.3 and 6.4. Hence, when $[-\pi, \pi]$ is considered as the continuous-time domain of an analog signal $f(x)$, the bi-infinite sequence $c_k = c_k(f)$ of the Fourier coefficients of $f(x)$ captures the frequency contents of the signal on the discrete-time domain $\ldots, -n, \ldots, n, \ldots$. In Sect. 6.3, the notion of Dirichlet's kernels, $D_n(x)$, is introduced as an ideal lowpass filter for removing the frequency contents of the signal $f(x)$ for all k outside the discrete-time interval $\{-n, \ldots, n\}$, while keeping the frequency contents in the finite discrete time interval $\{-n, \ldots, n\}$ intact. More precisely, with the filtering process defined by the (integral) convolution, the ideal lowpass filtering mentioned above is given by

$$(S_n f)(x) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t)D_n(x - t)dt,$$

where $(S_n f)(x)$ is the nth partial sum of the Fourier series of $f(x)$. The explicit formula

$$D_n(x) = \sin \left(n + \frac{1}{2} \right) x \Big/ \sin \frac{x}{2}$$

of nth order Dirichlet's kernel is derived in Sect. 6.3. In addition, since ideal lowpass filtering does not provide satisfactory time-domain localization (for the proof of convergence of the Fourier series), the notion of Fejér's kernels , defined by taking the averages of Dirichlet's kernels, $\sigma_n(x)$, is also introduced in Sect. 6.3, namely:

$$\sigma_n(x) = \frac{1}{n+1}\big(D_0(x) + \cdots + D_n(x)\big).$$

Indeed, the explicit formula

$$\sigma_n(x) = \frac{1}{n+1}\left(\sin\left(n+\frac{1}{2}\right) x \Big/ \sin\frac{x}{2}\right)^2$$

of the nth order Fejér's kernel derived in this section allows us to prove not only that the sequence $\{\sigma_n(x)\}$ constitutes a positive approximate identity, but also that the energy of the convolution of any continuous-time signal on $[-\pi, \pi]$ does not exceed (a constant multiple of) the energy of the signal itself. While the Fourier series of a function $f \in PC_{2\pi}^*$ does not necessarily converge to $f(x)$ at any point x, even if f does not have any jump discontinuities, it will be shown in Sect. 6.4 that, under the additional assumption of the existence of both one-sided derivatives of f at x, the sequence of partial sums $(S_n f)(x)$, formulated as convolution of f with Dirichlet's kernels, does converge to $f(x)$ at x. On the other hand, by applying the property of positive approximate identity of Fejér's kernels, it will also be shown that $f(x)$ can be recovered at any point x, where f is continuous, from the average (called Césaro means)

$$(C_n f)(x) = \frac{(S_0 f)(x) + \ldots + (S_n f)(x)}{n+1}$$

of the partial sums of the Fourier series $(Sf)(x)$ of $f(x)$, without the differentiability assumption. Consequently, in view of the "energy preservation" property of the convolution operation with Fejér's kernels, it follows that the Fourier series of any finite-energy function f converges to f in the energy (or L_2) norm.

Fourier series representations provide powerful mathematical tools in numerous areas of mathematical analysis and applications. Application of the discrete version to data compression, particularly to digital image compression, has been discussed in Chap. 5. In this final section of this chapter, we will apply the continuous version to solving linear partial differential equations (PDE's). Since the diffusion process is fundamental to data "noise removal", we emphasize our study to the diffusion PDE in this section. For the book to be self-contained, we include the study of the "method of separation of variables" to uncouple the PDE to the totality of an ordinary differential equation (ODE) in the time variable and an ODE (or PDE for higher spatial dimensions) defined only in the spatial domain. Separation of spatial variables to lower dimensions can be carried out analogously. For the heat diffusion PDE, a perfectly insulated spatial domain specifies the zero Neumann condition, and the initial heat content (or temperature) is represented by its Fourier series (for the rectangular

spatial domain). Our discussion for the one spatial-variable setting is extended to the setting of higher dimensional spatial "rectangular" regions. Application to solving the wave equation, particularly the problem of vibrating string (for one spatial variable with zero two-point boundary condition) and vibrating membrane (for zero boundary condition in two spatial variables) is left as exercises.

6.1 Fourier Series

Recall from Sect. 1.3 of Chap. 1 that $PC(J)$ denotes the set of piecewise continuous functions on the interval J. For $J = [a, b]$ where $-\infty < a < b < \infty$, we use the notation $PC[a, b]$ for $PC([a, b])$. Observe that there is no loss of generality by considering $[a, b] = [-\pi, \pi]$ by the simple change of variables:

$$x \longleftrightarrow \frac{2\pi}{b - a}(x - a) - \pi;$$

namely, the functions $\widetilde{f} \in PC[a, b]$ and $f \in PC[-\pi, \pi]$ are interchangeable by considering

$$f(x) = \widetilde{f}\left(a + \frac{b - a}{2\pi}(x + \pi)\right)$$

and

$$\widetilde{f}(x) = f\left(\frac{2\pi}{b - a}(x - a) - \pi\right).$$

Of course, each function $f \in PC[-\pi, \pi]$ can be extended to a 2π-periodic function.

Definition 1 $\boxed{\textbf{Periodic extension}}$ *Let $f(x)$ be a function defined on an interval $[-d, d]$. The $2d$-periodic extension $F(x)$ of $f(x)$ is a function on $\mathbb{R}$ defined by*

$$\begin{cases} F(x) = f(x), \ x \in (-d, d), \\ F(-d) = F(d) = \frac{1}{2}\big(f(d) + f(-d)\big), \end{cases}$$

and $F(x + 2\ell d) = F(x)$ for all $\ell \in \mathbb{Z}$. For convenience, we re-name $F(x)$ as the given function $f(x)$, to avoid unnecessary additional notation.

In this section, let $PC_{2\pi}^{\star} = PC^{\star}[-\pi, \pi]$ denote the inner-product space of such 2π-periodic piecewise continuous functions with inner product $\langle\!\langle , \rangle\!\rangle$ defined in Sect. 4.1 of Chap. 4 on p.172; that is,

$$\langle\!\langle f, g \rangle\!\rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x)\overline{g(x)}dx. \tag{6.1.1}$$

Observe that the only difference between this inner product $\langle\!\langle \cdot, \cdot \rangle\!\rangle$ and the inner product $\langle \cdot, \cdot \rangle$ for $\widetilde{L}_2[-\pi, \pi]$ in Sect. 1.3 of Chap. 1 on p.24 or $L_2[-\pi, \pi]$ is the additional normalization constant of $(2\pi)^{-1}$ in (6.1.1); that is,

$$\langle\!\langle f, g \rangle\!\rangle = \frac{1}{2\pi} \langle f, g \rangle.$$

In the following, let

$$e_k(x) = e^{ikx} = \cos kx + i \sin kx,$$

and let $\mathbb{V}_{2n+1}$ be the subspace of $PC_{2\pi}^\star$ defined by the algebraic span:

$$\mathbb{V}_{2n+1} = \text{span}\{e_k(x) : -n \le k \le n\}. \tag{6.1.2}$$

Then $\mathbb{V}_{2n+1}$ is a $(2n + 1)$-dimensional vector space spanned by $\{e_{-n}(x), \ldots, e_n(x)\}$. In fact, the set $\{e_{-n}(x), \ldots, e_n(x)\}$ is an orthonormal basis of $\mathbb{V}_{2n+1}$, since

$$\langle\!\langle e_j, e_k \rangle\!\rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} e_j(x)\overline{e_k(x)}dx = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{i(j-k)x}dx = \delta_{j-k}$$

for all $j, k = -n, \ldots, n$.

The continuous version of the discrete Fourier transformation (DFT) F_n, studied in Sect. 3.1 of Chap. 3, is the infinite sequence $\{c_k\}$ associated with a function $f(x)$ in $\mathbb{V} = PC_{2\pi}^\star$, defined, for each integer k, by

$$c_k = c_k(f) = \langle\!\langle f, e_k \rangle\!\rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x)\overline{e_k(x)}dx$$

or

$$c_k = c_k(f) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x)e^{-ikx}dx. \tag{6.1.3}$$

The analogy of the inverse DFT $(F_n)^{-1} = \frac{1}{n}(F_n)^*$ (see Definition 1 on p.174) is the following infinite series associated with $f(x)$, namely:

$$(Sf)(x) = \sum_{k=-\infty}^{\infty} c_k(f)e_k(x) = \sum_{k=-\infty}^{\infty} c_k e^{ikx}. \tag{6.1.4}$$

Definition 2 $\boxed{\text{Fourier series}}$ *Let* $f \in PC_{2\pi}^\star$. *Then the infinite series in* (6.1.4), *with* $c_k = c_k(f)$, *defined by* (6.1.3), *is called the Fourier series of* $f(x)$, *and* c_k *is called the* k*th Fourier coefficient of* $f(x)$.

Remark 1 Since the domain of the sequence $\{c_k(f)\}$ is the set $\mathbb{Z}$ of all integers, the "frequency-domain" of $f \in PC_{2\pi}^\star$ is $\mathbb{Z}$, while its "time-domain" is the inter-

val $[-\pi, \pi]$. The importance of the Fourier series representation is that the Fourier coefficient $c_k(f)$ reveals the kth frequency component of the function $f(x)$, and the Fourier series (or averages of its partial sums) can be used to recover $f(x)$ from its frequency content $\{c_k(f)\}$. ∎

Example 1 Let $f_1(x)$ be defined by

$$f_1(x) = \begin{cases} 1, & \text{for } 0 \le x \le \pi, \\ -1, & \text{for } -\pi \le x < 0, \end{cases}$$

and extended 2π-periodically to all $x \in \mathbb{R}$ (see Definition 1). Compute the Fourier coefficients and Fourier series of $f_1(x)$.

Solution By (6.1.3), we have

$$\begin{aligned} c_k &= c_k(f_1) = \frac{1}{2\pi} \left\{ \int_{-\pi}^{0} (-1)e^{-ikx} dx + \int_{0}^{\pi} e^{-ikx} dx \right\} \\ &= \frac{1}{2\pi} \int_{0}^{\pi} \left(-e^{ikx} + e^{-ikx} \right) dx \\ &= \frac{-i}{\pi} \int_{0}^{\pi} \sin kx \, dx. \end{aligned}$$

Hence, $c_0 = 0$ and for $k \ne 0$,

$$c_k = \left[\frac{i}{\pi} \frac{\cos kx}{k} \right]_0^{\pi} = \frac{i}{\pi k} \left((-1)^k - 1 \right).$$

In other words, the Fourier coefficients of $f_1(x)$ are given by

$$c_{2\ell} = 0, \quad c_{2\ell+1} = \frac{-2i}{\pi(2\ell+1)}, \text{ for all } \ell = 0, \pm 1, \pm 2, \dots \qquad (6.1.5)$$

(where we consider $k = 2\ell$ and $k = 2\ell + 1$ separately), and the Fourier series of $f_1(x)$ is given by

$$\begin{aligned} (Sf_1)(x) &= \frac{2}{i\pi} \sum_{\ell=-\infty}^{\infty} \frac{e^{i(2\ell+1)x}}{2\ell+1} \\ &= \frac{2}{i\pi} \left(\sum_{\ell=0}^{\infty} \frac{e^{i(2\ell+1)x}}{2\ell+1} + \sum_{\ell=-\infty}^{-1} \frac{e^{i(2\ell+1)x}}{2\ell+1} \right) \end{aligned}$$

$$= \frac{2}{i\pi} \left\{ (e^{ix} - e^{-ix}) + \frac{e^{i3x} - e^{-i3x}}{3} + \frac{e^{i5x} - e^{-i5x}}{5} + \cdots \right\}$$

$$= \frac{4}{\pi} \sum_{k=0}^{\infty} \frac{\sin(2k+1)x}{2k+1}.$$

$\blacksquare$

Example 2 Let $g_1(x)$ be defined on $[-\pi, \pi]$ by

$$g_1(x) = \begin{cases} 1, & \text{for } |x| \leq \frac{\pi}{2}, \\ 0, & \text{for } \frac{\pi}{2} < |x| \leq \pi, \end{cases}$$

and extended periodically to $\mathbb{R}$. Compute the Fourier coefficients and Fourier series of $g_1(x)$.

Solution By (6.1.3), we have

$$c_k = c_k(g_1) = \frac{1}{2\pi} \int_{-\pi}^{\pi} g_1(x) e^{-ikx} dx = \frac{1}{2\pi} \int_{-\pi/2}^{\pi/2} e^{-ikx} dx,$$

so that $c_0 = \frac{1}{2}$ and for $k \neq 0$,

$$\begin{aligned} c_k &= \frac{1}{2\pi} \int_{-\pi/2}^{\pi/2} \left(\cos kx - i \sin kx \right) dx \\ &= \frac{1}{2\pi} \int_{-\pi/2}^{\pi/2} \cos kx \, dx = \frac{1}{\pi} \int_{0}^{\pi/2} \cos kx \, dx \\ &= \left[\frac{1}{\pi} \frac{\sin kx}{k} \right]_{0}^{\pi/2} = \frac{1}{\pi} \frac{\sin k\pi/2}{k}. \end{aligned}$$

Again, since $c_k = 0$ for all even $k \neq 0$, we only consider odd $k = 2\ell + 1$, for which

$$\sin \frac{(2\ell + 1)\pi}{2} = \sin \left(\ell\pi + \frac{\pi}{2} \right) = (-1)^{\ell}.$$

Therefore, we have $c_0 = \frac{1}{2}$, $c_{2\ell} = 0$ for all $\ell = \pm 1, \pm 2, \ldots$, and

$$c_{2\ell+1} = \frac{(-1)^{\ell}}{(2\ell + 1)\pi}, \quad \ell = 0, \pm 1, \pm 2, \ldots;$$

and the Fourier series of $g_1(x)$ is given by

$$(Sg_1)(x) = \frac{1}{2} + \sum_{\ell=-\infty}^{\infty} \frac{(-1)^\ell}{(2\ell+1)\pi} e^{i(2\ell+1)x}$$

$$= \frac{1}{2} + \frac{1}{\pi}\left\{(e^{ix} + e^{-ix}) - \frac{e^{i3x} + e^{-i3x}}{3} + \cdots\right\}$$

$$= \frac{1}{2} + \frac{2}{\pi}\sum_{k=0}^{\infty} (-1)^k \frac{\cos(2k+1)x}{2k+1}.$$

∎

So, what does the Fourier series $(Sf)(x)$ have to do with the function $f \in PC_{2\pi}^\star$?
To answer this question, we consider the nth partial sum

$$(S_n f)(x) = \sum_{k=-n}^{n} c_k(f)e_k(x) = \sum_{k=-n}^{n} c_k e^{ikx} \tag{6.1.6}$$

of the Fourier series $(Sf)(x)$, defined in (6.1.4). Observe that $(S_n f)(x)$ is the orthog-
onal projection of f from $PC_{2\pi}^\star$ onto $\mathbb{V}_{2n+1}$, defined by (6.1.2) with the inner product
$\langle\!\langle \cdot, \cdot \rangle\!\rangle$ of $PC_{2\pi}^\star$ (see definition of orthogonal projection on p.41 in Sect. 1.4 of Chap. 1).
Thus, by Theorem 2 on p.43, we have

$$\frac{1}{2\pi}\|f - S_n f\|_2^2 \leq \frac{1}{2\pi}\|f - g\|_2^2, \text{ for all } g \in \mathbb{V}_{2n+1},$$

and furthermore, by Theorem 3 on p.44, we have

$$\frac{1}{2\pi}\|f - S_n f\|_2^2 = \frac{1}{2\pi}\|f\|_2^2 - \sum_{k=-n}^{n} |c_k|^2. \tag{6.1.7}$$

This establishes the first statement in the following theorem. The second statement
in the theorem will be proved in Sect. 6.4 (see Remark 1 on p.304).

Theorem 1 $\boxed{\textbf{Best } L_2\textbf{-approximation and convergence}}$ *Let $S_n f$ be the nth partial
sum of the Fourier series of $f(x) \in PC_{2\pi}^\star$. Then*

$$\|f - S_n f\|_2 = \inf\{\|f - g\|_2 : g \in \mathbb{V}_{2n+1}\}, \tag{6.1.8}$$

and the sequence $\{S_n f\}$ converges to f in $L_2[-\pi, \pi]$, namely:

$$\|f - S_n f\|_2 \to 0, \text{ as } n \to \infty, \tag{6.1.9}$$

where $\|\cdot\|_2$ denotes the $L_2[-\pi, \pi]$-norm.

As a consequence of Theorem 1, we have the following result.

Theorem 2 $\boxed{\textbf{Parseval's identity}}$ *Let $f \in PC_{2\pi}^\star$ and $\{c_k\}$ be the sequence of Fourier coefficients of f as defined in* (6.1.4). *Then*

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} |f(x)|^2 dx = \sum_{k=-\infty}^{\infty} |c_k|^2. \tag{6.1.10}$$

This identity assures that the "energy" of the sequence $\{c_k\}$ of the Fourier coefficients of $f \in PC_{2\pi}^\star$ is governed by the energy of f. The identity (6.1.10), which is valid for all $f \in PC_{2\pi}^\star$, is called "Parseval's identity".

The proof of Theorem 2 follows from (6.1.7) and Theorem 1, since

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} |f(x)|^2 dx - \sum_{k=-n}^{n} |c_k|^2 = \frac{1}{2\pi} \|f - S_n f\|_2^2 \to 0$$

for $n \to 0$.   ∎

It also follows from Theorem 1 that the family $\{e_k(x)\}, k = 0, \pm 1, \pm 2, \ldots$, is an orthonormal basis of the inner-product space $PC[-\pi, \pi]$ with inner product $\langle\!\langle \cdot, \cdot \rangle\!\rangle$. Since $PC[-\pi, \pi]$ is dense in $L_2[-\pi, \pi]$, namely any function $f \in L_2[-\pi, \pi]$ can be approximated as closely as desired in the $L_2[-\pi, \pi]$-norm by functions in $PC[-\pi, \pi]$, we may conclude that $\{e_k(x)\}, k = 0, \pm 1, \pm 2, \ldots$, is an orthogonal basis of $L_2[-\pi, \pi]$. That is, any $f \in L_2[-\pi, \pi]$ can be written as

$$f(x) = \sum_{k=-\infty}^{\infty} c_k(f) e^{ikx},$$

where the infinite series converges in $L_2[-\pi, \pi]$, and where

$$c_k(f) = \langle\!\langle f, e_k \rangle\!\rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) e^{-ikx} dx,$$

which is the same integral as the one defined in (6.1.3). The above discussion is also valid when $[-\pi, \pi]$ is replaced by $[-d, d]$ with $d > 0$. More precisely, we have the following theorem.

Theorem 3 $\boxed{\textbf{Fourier series of functions in } L_2[-d, d]}$ *For $d > 0$, the family $\{e^{ik\pi x/d}\}, k = 0, \pm 1, \pm 2, \ldots$, is an orthonormal basis of $L_2[-d, d]$ with inner product $\frac{1}{2d}\langle \cdot, \cdot \rangle_{L_2[-d,d]}$. Consequently, any function f in $L_2[-d, d]$ can be represented by its Fourier series, namely:*

$$f(x) = (Sf)(x) = \sum_{k=-\infty}^{\infty} c_k(f) e^{ik\pi x/d}, \tag{6.1.11}$$

which converges to f *in* $L_2[-d, d]$, *where*

$$c_k(f) = \frac{1}{2d} \int_{-d}^{d} f(x)e^{-ik\pi x/d} dx.$$

Furthermore, the Fourier coefficients $c_k = c_k(f)$, $k = 0, \pm 1, \pm 2, \ldots$, *satisfy Parseval's identity, namely:*

$$\frac{1}{2d} \int_{-d}^{d} |f(x)|^2 dx = \sum_{k=-\infty}^{\infty} |c_k|^2, \quad \text{for all } f \in L_2[-d, d].$$

In the rest of this section, we apply Parseval's identity (6.1.10) for some suitably chosen functions to compute the precise values of certain interesting series.

Example 3 Apply Parseval's identity (6.1.10) to the function $f_1(x)$ defined in Example 1 to compute the precise value of the infinite series

$$\sum_{k=1}^{\infty} \frac{1}{k^2}.$$

Solution Since $\| f_1 \|_2^2 = 2\pi$, it follows from (6.1.5) that

$$\begin{aligned}
1 &= \sum_{\ell=-\infty}^{\infty} \left| \frac{-2i}{\pi(2\ell+1)} \right|^2 \\
&= \frac{4}{\pi^2} \left(\sum_{\ell=0}^{\infty} \frac{1}{(2\ell+1)^2} + \sum_{\ell=-\infty}^{-1} \frac{1}{(2\ell+1)^2} \right) \\
&= \frac{4}{\pi^2} \left(\sum_{\ell=0}^{\infty} \frac{1}{(2\ell+1)^2} + \sum_{\ell=1}^{\infty} \frac{1}{(2\ell-1)^2} \right) \\
&= \frac{4}{\pi^2} \left(\sum_{\ell=0}^{\infty} \frac{1}{(2\ell+1)^2} + \sum_{\ell=0}^{\infty} \frac{1}{(2\ell+1)^2} \right) \\
&= \frac{8}{\pi^2} \sum_{\ell=0}^{\infty} \frac{1}{(2\ell+1)^2},
\end{aligned}$$

or

$$\sum_{\ell=0}^{\infty} \frac{1}{(2\ell+1)^2} = \frac{\pi^2}{8}. \tag{6.1.12}$$

To complete the solution of Example 3, we simply partition the sum over k into two sums, with one over even k's and the other over odd k's, namely:

$$\sum_{k=1}^{\infty} \frac{1}{k^2} = \sum_{\ell=1}^{\infty} \frac{1}{(2\ell)^2} + \sum_{\ell=0}^{\infty} \frac{1}{(2\ell+1)^2}$$

$$= \frac{1}{4} \sum_{\ell=1}^{\infty} \frac{1}{\ell^2} + \frac{\pi^2}{8}, \qquad (6.1.13)$$

by applying (6.1.12), so that

$$\sum_{k=1}^{\infty} \frac{1}{k^2} = \left(1 - \frac{1}{4}\right)^{-1} \frac{\pi^2}{8} = \frac{4}{3} \frac{\pi^2}{8} = \frac{\pi^2}{6}. \qquad (6.1.14)$$

∎

Remark 2 Example 3 is called the "Basel problem", posed by P. Mengoli in the year 1644 and solved by L. Euler in 1735, before the introduction of Fourier series by J. Fourier in the early 1800's. In fact, Euler has derived the exact formula

$$\zeta(2n) = \sum_{k=1}^{\infty} \frac{1}{k^{2n}} = \frac{(-1)^{n-1} 2^{2n-1} \pi^{2n}}{(2n)!} b_{2n}, \qquad (6.1.15)$$

for all $n = 1, 2, \ldots$, where the b_{2n}'s are the Bernoulli numbers, whose exact values can be easily computed recursively. The function $\zeta(z)$ defined for the complex variable z is the famous "Riemann zeta function". The first three Bernoulli numbers b_{2n} are:

$$b_2 = \frac{1}{6}, \quad b_4 = \frac{-1}{30}, \quad b_6 = \frac{1}{42}. \qquad (6.1.16)$$

On the other hand, the problem of finding the exact values of

$$\zeta(2n + 1) = \sum_{k=1}^{\infty} \frac{1}{k^{2n+1}}, \quad n = 1, 2, \ldots$$

remains unsolved, although Euler was able to compute at least $\zeta(3)$ and $\zeta(5)$ fairly accurately, of course without the computer (in the eighteenth century). ∎

Remark 3 Putting the value of b_2 in (6.1.16) into the formula (6.1.15) yields $\zeta(2) = \pi^2/6$, which is precisely the solution (6.1.14) of Example 3; where the Fourier series of the function $f_1(x)$ in Example 1 is used in applying Parseval's identity. In the following examples, we further demonstrate the power of the Fourier methods for computing $\zeta(2n)$ without using Euler's formula (6.1.15) and Bernoulli numbers b_{2n}. The idea is to find functions $f_n(x)$ that satisfy $f_n(-\pi) = f_n(\pi)$ such that $f_n'(x)$ can be expressed in terms of $f_{n-1}(x)$, so that the Fourier coefficients $c_k(f_n)$ can be computed from $c_k(f_{n-1})$ simply by integration by parts. ∎

In the following two examples, we illustrate the idea proposed in Remark 3, but leave the detailed calculations as exercises.

Example 4 Apply Parseval's identity to compute $\zeta(4)$ by selecting an appropriate function $f_2 \in PC_{2\pi}^\star$. Verify the answer with Euler's formula (6.1.15) and the Bernoulli number b_4 in (6.1.16).

Solution Following the idea introduced in Remark 3, we set $f_2(x) = |x|$ for $x \in [-\pi, \pi]$ and extend this function from $[-\pi, \pi]$ to $\mathbb{R}$ by setting $f_2(x) = f_2(x + 2\pi)$, so that $f_2 \in PC_{2\pi}^\star$. Observe that $f_2'(x) = f_1(x)$ in Example 1. Hence, by integration by parts, we have, for $k \neq 0$,

$$
\begin{aligned}
c_k(f_2) &= \frac{1}{2\pi}\left[f_2(x)\frac{e^{-ikx}}{-ik}\right]_{-\pi}^\pi - \frac{1}{2\pi}\int_{-\pi}^\pi f_1(x)\frac{e^{-ikx}}{-ik}\,dx \\
&= \frac{-i}{k}c_k(f_1),
\end{aligned}
\tag{6.1.17}
$$

where the first term vanishes due to the fact that $f_2(-\pi) = f_2(\pi)$. Since $c_0(f_2) = \frac{\pi}{2}$ and $c_{2\ell}(f_1) = 0$ for all $\ell \neq 0$, it follows from (6.1.5) and (6.1.17) that

$$
\begin{aligned}
\sum_{k=-\infty}^\infty |c_k(f_2)|^2 &= \frac{\pi^2}{4} + \sum_{\ell=-\infty}^\infty \frac{4}{\pi^2(2\ell+1)^4} \\
&= \frac{\pi^2}{4} + \frac{8}{\pi^2}\sum_{\ell=0}^\infty \frac{1}{(2\ell+1)^4} \\
&= \frac{\pi^2}{4} + \left(1 - \frac{1}{2^4}\right)^{-1}\frac{8}{\pi^2}\sum_{k=1}^\infty \frac{1}{k^4},
\end{aligned}
$$

where the trick introduced in the derivation of (6.1.13) is applied to arrive at the quantity in the last line. Hence, by Parseval's identity, we have

$$
\begin{aligned}
\sum_{k=1}^\infty \frac{1}{k^4} &= \left(\frac{1}{2\pi}\|f_2\|_2^2 - \frac{\pi^2}{4}\right)\frac{2^4}{2^4-1}\frac{\pi^2}{8} \\
&= \left(\frac{\pi^2}{3} - \frac{\pi^2}{4}\right)\frac{2^4}{2^4-1}\frac{\pi^2}{2^3} = \frac{\pi^4}{90},
\end{aligned}
\tag{6.1.18}
$$

which exactly matches the value of $\zeta(4)$ in (6.1.15) for $n = 2$ by using the value of b_4 in (6.1.16) (see Exercise 6). ∎

Example 5 Follow the instruction in Example 4 to compute $\zeta(6)$ without using Euler's formula (6.1.15).

Solution Following the idea introduced in Remark 3, we consider the odd function extension of

$$f_3(x) = \frac{\pi^2}{8} - \frac{1}{2}\left(x - \frac{\pi}{2}\right)^2, \quad 0 \le x \le \pi, \qquad (6.1.19)$$

by setting $f_3(x) = -f_3(-x)$ for $-\pi \le x < 0$. Then extend $f_3(x)$ from $[-\pi, \pi]$ periodically to $f_3 \in PC_{2\pi}^\star$. Observe that

$$f_3'(x) = \frac{\pi}{2} - f_2(x),$$

$f_3(-\pi) = f_3(\pi) = 0$, and that

$$\frac{1}{2\pi}\|f_3\|_2^2 = \frac{\pi^4}{120} \qquad (6.1.20)$$

(see Exercise 8 for the computation of (6.1.20)). Hence, by applying (6.1.17) and by following the same method of derivation for (6.1.18), we have

$$\sum_{k=1}^{\infty} \frac{1}{k^6} = \frac{2^6}{2^6 - 1} \frac{\pi^2}{8} \frac{1}{2\pi}\|f_3\|_2^2$$

$$= \frac{8\pi^2}{63} \cdot \frac{\pi^4}{120} = \frac{\pi^6}{945}, \qquad (6.1.21)$$

which exactly matches the value of $\zeta(6)$ in (6.1.15) for $n = 3$ (where the value of b_6 in (6.1.16) is used). ∎

Exercises

Exercise 1 Let $f \in PC_{2\pi}^\star$. Show that

$$\int_{a-\pi}^{a+\pi} f(x)dx = \int_{-\pi}^{\pi} f(x)\, dx$$

for all real numbers a.

Exercise 2 Compute the Fourier series $(Sf)(x)$ of the "saw-tooth" function $f(x)$ defined on $[-\pi, \pi]$ by $f(x) = x$ and extended periodically to $\mathbb{R}$.

Exercise 3 Let $g(x) = \pi + x$ for $-\pi \le x \le 0$ and $g(x) = \pi - x$ for $0 < x \le \pi$. Compute the Fourier series $(Sg)(x)$ of $g(x)$.

Exercise 4 Compute the Fourier series (6.1.4) of each of the following functions:

(a) $g_1(x) = \cos^2 x$,
(b) $g_2(x) = 1 + 2\sin^2 x$,
(c) $g_3(x) = \sin^3 x$,
(c) $g_4(x) = g_1(x) - g_2(x) + g_3(x)$.

Exercise 5 Let $d > 0$. Verify that the functions $\frac{1}{\sqrt{2d}}e^{ik\pi x/d}$, $k = 0, \pm 1, \pm 2, \ldots$, constitute an orthonormal family in $L_2[-d, d]$.

Exercise 6 Fill in the computational details in (6.1.18) by computing $\|f_2\|_2^2$ and splitting the infinite sum into the sum over all even integers and the sum over all odd integers.

Exercise 7 Verify that $f_3(x)$, as defined in (6.1.19), satisfies $f_3(-\pi) = f_3(\pi)$ and (6.1.20).

Exercise 8 As a continuation of Exercises 6–7, fill in the computational details to derive the formula (6.1.21).

Exercise 9 Apply Parseval's identity to show that the Fourier series of any $f \in PC_{2\pi}^\star$ determines f uniquely, in the sense that if two functions in $PC_{2\pi}^\star$ have the same Fourier series, then these two functions are equal almost everywhere (see the notation of "almost everywhere" in Sect. 1.1).

Exercise 10 Show that the sequence $\{c_k\}$ of Fourier coefficients of any $f \in L_2[-d, d]$ converges to 0, as $k \to \pm\infty$.

Hint: Apply Parseval's identity in Theorem 3.

6.2 Fourier Series in Cosines and Sines

For various important reasons, including the study of (Fourier) cosine series and facilitation of implementation, we convert the Fourier series expansion of functions $f \in PC_{2\pi}^\star$ to (Fourier) trigonometric series in terms of cosines and sines by getting rid of the imaginary unit, i, in Euler's formula, as follows.

Theorem 1 | **Fourier series in cosines and sines** | *The Fourier series Sf of $f \in PC_{2\pi}^\star$ in (6.1.4) on p.267 can be re-formulated as*

$$(Sf)(x) = \frac{a_0}{2} + \sum_{k=1}^{\infty} \left(a_k \cos kx + b_k \sin kx \right), \qquad (6.2.1)$$

where

$$a_k = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos kx dx, k = 0, 1, 2, \ldots,$$

and

$$b_k = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin kx dx, k = 1, 2, \ldots.$$

Furthermore, the nth partial sum of $(Sf)(x)$ in (6.1.6) on p.270 can be written as

$$(S_n f)(x) = \frac{a_0}{2} + \sum_{k=1}^{n} \left(a_k \cos kx + b_k \sin kx \right), \qquad (6.2.2)$$

and converges to $f(x)$ in $L_2[-\pi, \pi]$, in the sense that

$$\| f - S_n f \|_2 \to 0, \text{ as } n \to \infty. \qquad (6.2.3)$$

In addition, the Fourier cosine and sine coefficients in (6.2.1) satisfy Parseval's identity; that is,

$$\frac{|a_0|^2}{4} + \frac{1}{2} \sum_{k=1}^{\infty} \left(|a_k|^2 + |b_k|^2 \right) = \frac{1}{2\pi} \int_{-\pi}^{\pi} |f(x)|^2 dx, \qquad (6.2.4)$$

for all $f \in PC_{2\pi}^\star$.

Proof Observe that

$$\frac{a_0}{2} = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x)dx = c_0,$$

where c_0 is defined in (6.1.3) on p.267 for $k = 0$. For $k = 1, 2, \ldots$,

$$c_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x)e^{-ikx} dx = \frac{1}{2}(a_k - ib_k);$$

$$c_{-k} = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x)e^{ikx} dx = \frac{1}{2}(a_k + ib_k),$$

so that

$$
\begin{aligned}
(Sf)(x) &= c_0 + \sum_{k=1}^{\infty} c_k e^{ikx} + \sum_{k=-\infty}^{-1} c_k e^{ikx} \\
&= c_0 + \sum_{k=1}^{\infty} c_k e^{ikx} + \sum_{k=1}^{\infty} c_{-k} e^{-ikx} \\
&= \frac{a_0}{2} + \sum_{k=1}^{\infty} \frac{1}{2}(a_k - ib_k)e^{ikx} + \frac{1}{2}(a_k + ib_k)e^{-ikx} \\
&= \frac{a_0}{2} + \sum_{k=1}^{\infty} \frac{1}{2}a_k(e^{ikx} + e^{-ikx}) + \frac{1}{2}(-ib_k)(e^{ikx} - e^{-ikx}) \\
&= \frac{a_0}{2} + \sum_{k=1}^{\infty} a_k \cos kx + b_k \sin kx.
\end{aligned}
$$

This proves (6.2.1).

Following the above derivation, we also have

$$(S_n f)(x) = \frac{a_0}{2} + \sum_{k=1}^{n} (a_k \cos kx + b_k \sin kx) = \sum_{k=-n}^{n} c_k e^{ikx}.$$

Thus, the convergence property in (6.2.3) follows from (6.1.9) on p.270.

Furthermore, Parseval's identity (6.2.4) follows from (6.1.10) on p.271, since $c_0 = a_0/2$ and

$$\begin{aligned}|c_k|^2 + |c_{-k}|^2 &= \frac{1}{4}|a_k - ib_k|^2 + \frac{1}{4}|a_k + ib_k|^2 \\ &= \frac{1}{2}\left(|a_k|^2 + |b_k|^2\right), k = 1, 2, \ldots,\end{aligned}$$

where the fact that

$$|a - ib|^2 + |a + ib|^2 = 2\left(|a|^2 + |b|^2\right), \tag{6.2.5}$$

which is valid for all complex numbers a and b (see Exercise 2), is applied. ■

Since $PC[-\pi, \pi]$ is dense in $L_2[-\pi, \pi]$, it follows from Theorem 1 on p.270 that the orthonormal family $\{\frac{1}{\sqrt{2\pi}} e^{ikx} : k = 0, \pm 1, \ldots\}$ is an orthonormal basis of $L_2[-\pi, \pi]$. As an alternative, it follows from Theorem 1 that the orthonormal family

$$\left\{ \frac{1}{\sqrt{2\pi}}, \frac{1}{\sqrt{\pi}} \cos kx, \frac{1}{\sqrt{\pi}} \sin kx : \quad k = 1, 2, \ldots \right\}$$

of real-valued functions is another orthonormal basis of $L_2[-\pi, \pi]$. In general, we have the following theorem.

Theorem 2 | **Fourier cosine and sineseries for $L_2[-d, d]$** | *Let $d > 0$. Then the family*

$$W = \left\{ \frac{1}{\sqrt{2d}}, \frac{1}{\sqrt{d}} \cos \frac{k\pi x}{d}, \frac{1}{\sqrt{d}} \sin \frac{k\pi x}{d} : k = 1, 2, \ldots \right\} \tag{6.2.6}$$

is an orthonormal basis of $L_2[-d, d]$. Consequently, any f in $L_2[-d, d]$ can be represented by its Fourier series in cosines and sines, namely:

$$f(x) = (Sf)(x) = \frac{a_0}{2} + \sum_{k=1}^{\infty} \left(a_k \cos \frac{k\pi x}{d} + b_k \sin \frac{k\pi x}{d} \right), \tag{6.2.7}$$

which converges to f in $L_2[-d, d]$, where

$$a_k = \frac{1}{d} \int_{-d}^{d} f(x) \cos \frac{k\pi x}{d} dx, \; for \; k = 0, 1, 2, \ldots, \tag{6.2.8}$$

and

$$b_k = \frac{1}{d} \int_{-d}^{d} f(x) \sin \frac{k\pi x}{d} dx, \; for \; k = 1, 2, \ldots. \tag{6.2.9}$$

Furthermore, the Fourier cosine and sine coefficients, defined in (6.2.8)–(6.2.9), *satisfy Parseval's identity, namely:*

$$\frac{|a_0|^2}{4} + \frac{1}{2} \sum_{k=1}^{\infty} \left(|a_k|^2 + |b_k|^2 \right) = \frac{1}{2d} \int_{-d}^{d} |f(x)|^2 dx.$$

Example 1 Let $f(x) = 1$ for $-d \le x \le 0$, and $f(x) = 0$ for $0 < x \le d$. Compute the Fourier cosine and sine series Sf of f in (6.2.7).

Solution For $k = 0$,

$$a_0 = \frac{1}{d} \int_{-d}^{d} f(x) dx = \frac{1}{d} \int_{-d}^{0} 1 dx = \frac{d}{d} = 1,$$

and for $k \ge 1$,

$$a_k = \frac{1}{d} \int_{-d}^{d} f(x) \cos \frac{k\pi x}{d} dx = \frac{1}{d} \int_{-d}^{0} 1 \cdot \cos \frac{k\pi x}{d} dx$$

$$= \left[\frac{1}{d} \frac{\sin \frac{k\pi x}{d}}{\frac{k\pi}{d}} \right]_{-d}^{0} = \frac{1}{k\pi} \left(\sin 0 - \sin(-k\pi) \right) = 0;$$

$$b_k = \frac{1}{d} \int_{-d}^{d} f(x) \sin \frac{k\pi x}{d} dx = \frac{1}{d} \int_{-d}^{0} 1 \cdot \sin \frac{k\pi x}{d} dx$$

$$= \left[-\frac{1}{d} \frac{\cos \frac{k\pi x}{d}}{\frac{n\pi}{d}} \right]_{-d}^{0} = -\frac{1}{k\pi} \left(\cos 0 - \cos(-k\pi) \right)$$

$$= -\frac{1}{k\pi} \left(1 - (-1)^k \right)$$

Thus, the Fourier cosine and sine series of f is given by

$$(Sf)(x) = \frac{1}{2} - \sum_{k=1}^{\infty} \frac{1}{k\pi} (1 - (-1)^k) \sin \frac{k\pi x}{d}$$

$$= \frac{1}{2} - \sum_{n=0}^{\infty} \frac{2}{(2n+1)\pi} \sin \frac{(2n+1)\pi x}{d}.$$

■

For the Fourier series expansion $(Sf)(x)$ in terms of both cosines and sines in (6.2.1) of Theorem 1 or in (6.2.7) in Theorem 2, we observe that the computational complexity can be reduced by eliminating either the sine series or the cosine series. Recall from the first paragraph of Sect. 6.1 that any function $f(x)$, defined on an interval $[a, b]$, can be treated as a function defined on the interval $[0, d]$, by the change of variables $t = \frac{d}{b-a}(x - a)$. We then proceed by considering the Fourier trigonometric series of $\widetilde{f}(t)$ (with $\widetilde{f}(t) = f(x)$) on $[0, d]$ (instead of $f(x)$ on $[a, b]$), while recovering both $f(x)$ and its Fourier cosine and sine series by the change of variable $x = \frac{b-a}{d}t + a$. Hence, without loss of generality, we only consider functions $f(x)$ in $L_2[0, d]$, $d > 0$. For $f \in L_2[0, d]$, we may extend f to an even function f_e on $[-d, d]$ by setting

$$f_e(x) = \begin{cases} f(x), & \text{for } 0 \le x \le d, \\ f(-x), & \text{for } -d \le x < 0. \end{cases}$$

Then, since $\sin(\frac{k\pi x}{d})$ is an odd function, we have $b_k(f_e) = 0$ in (6.2.9), so that

$$(Sf_e)(x) = \frac{a_0(f_e)}{2} + \sum_{k=1}^{\infty} a_k(f_e) \cos \frac{k\pi x}{d}, \qquad (6.2.10)$$

where

$$a_k(f_e) = \frac{1}{d} \int_{-d}^{d} f_e(x) \cos \frac{k\pi x}{d} dx = \frac{2}{d} \int_{0}^{d} f(x) \cos \frac{k\pi x}{d} dx$$

for $k = 0, 1, 2, \ldots$. Let

$$(S_n f_e)(x) = \frac{a_0(f_e)}{2} + \sum_{k=1}^{n} a_k(f_e) \cos \frac{k\pi x}{d}$$

denote the nth partial sum of $(Sf_e)(x)$. Then it follows from Theorem 2 that $(S_n f_e)(x)$ converges to f_e in $L_2[-d, d]$. Thus, when restricted to $[0, d]$, $(S_n f_e)(x)$ converges to f in $L_2[0, d]$. The series in (6.2.10) is called the **cosine series** of f, and is denoted by $S^c f$. For functions defined on $[0, d]$, the collection

$$W_1 = \left\{ \frac{1}{\sqrt{d}}, \sqrt{\frac{2}{d}} \cos \frac{k\pi x}{d} : k = 1, 2, \ldots \right\} \qquad (6.2.11)$$

is an orthonormal family in $L_2[0, d]$ (see Exercise 5). Thus, W_1 is an orthonormal basis of $L_2[0, d]$. To summarize, we have the following theorem.

Theorem 3 | **Fourier cosine series** | *Let $d > 0$. Then the set W_1 of cosine functions defined by (6.2.11) is an orthonormal basis of $L_2[0, d]$. Thus, any function $f \in L_2[0, d]$ can be expanded as its cosine series, namely:*

$$f(x) = (S^c f)(x) = \frac{a_0}{2} + \sum_{k=1}^{\infty} a_k \cos \frac{k\pi x}{d}, \qquad (6.2.12)$$

which converges to $f(x)$ in $L_2[0, d]$, where the cosine coefficients are given by

$$a_k = \frac{2}{d} \int_0^d f(x) \cos \frac{k\pi x}{d} dx, \, k = 0, 1, \ldots. \qquad (6.2.13)$$

Example 2 Let $f(x) = 1 + 2\cos^2 x$, $0 \le x \le \pi$. Find the cosine series $(S^c f)$ of f in (6.2.12) with $d = \pi$.

Solution By the trigonometric identity $\cos^2 x = \frac{1}{2}(1 + \cos 2x)$, we may write $f(x)$ as

$$f(x) = 1 + 2 \cdot \frac{1}{2}(1 + \cos 2x) = 2 + \cos 2x.$$

Hence, applying (6.2.13), we can compute

$$a_0 = \frac{2}{\pi} \int_0^\pi (2 + \cos 2x) \, dx = 4;$$

$$a_2 = \frac{2}{\pi} \int_0^\pi (2 + \cos 2x) \cos 2x dx$$

$$= \frac{2}{\pi} \int_0^\pi \cos^2 2x dx = \frac{2}{\pi} \frac{\pi}{2} = 1,$$

and for $k = 3, 4, \ldots$, $a_k = 0$, since $\cos kx$ is orthogonal to 1 and $\cos 2x$. Thus, by (6.2.12), the cosine series $S^c f$ of $f(x) = 1 + 2\cos^2 x$ is

$$(S^c f)(x) = \frac{a_0}{2} + a_2 \cos 2x = 2 + \cos 2x.$$

Observe that there is really no need to compute a_k, since the cosine polynomial $2 + \cos 2x$ is already a cosine series of $\cos kx$, $k = 0, 1, \ldots$. ∎

In general, for integer powers of the cosine and sine functions, we may apply Euler's formula to avoid integration.

Example 3 Let $f(x) = \cos^5 x$, $0 \le x \le \pi$. Find the cosine series $(S^c f)$ of f in (6.2.12) with $d = \pi$.

Solution By Euler's formula, we have

$$\cos^5 x = \left(\frac{e^{ix} + e^{-ix}}{2} \right)^5$$

$$= \frac{1}{32} \left(e^{i5x} + 5e^{i3x} + 10e^{ix} + 10e^{-ix} + 5e^{-i3x} + e^{-i5x} \right)$$

$$= \frac{1}{32} \left(e^{i5x} + e^{-i5x} + 5(e^{i3x} + e^{-i3x}) + 10(e^{ix} + e^{-ix}) \right)$$

$$= \frac{1}{16} \left(\cos 5x + 5 \cos 3x + 10 \cos x \right).$$

Since the above trigonometric polynomial is already a "series" in terms of cosine basis functions $\cos kx,\, k = 0, 1, \ldots$, we know that its cosine series is itself. Thus, the cosine series $S^c f$ of $f(x) = \cos^5 x$ is given by

$$(S^c f)(x) = \frac{5}{8} \cos x + \frac{5}{16} \cos 3x + \frac{1}{16} \cos 5x,$$

where the terms are arranged in increasing order of "frequencies".   ■

Similarly, a function $f \in L_2[0, d]$ can be expanded as a sine series. The trick is to consider the odd extension f_o of f to $[-d, d]$, namely:

$$f_o(x) = \begin{cases} f(x), & \text{for } 0 < x \le d, \\ -f(-x), & \text{for } -d \le x < 0, \\ 0, & \text{for } x = 0. \end{cases}$$

Then, since $\cos \frac{k\pi x}{d}$ is an even function for each $k = 0, 1, \ldots$, we have $a_k(f_o) = 0$ in (6.2.8), and

$$b_k(f_o) = \frac{1}{d} \int_{-d}^{d} f_o(x) \sin \frac{k\pi x}{d} dx = \frac{2}{d} \int_{0}^{d} f(x) \sin \frac{k\pi x}{d} dx$$

for $k = 1, 2, \ldots$. Thus, the Fourier series of f_o is given by

$$(Sf_o)(x) = \sum_{k=1}^{\infty} b_k(f_o) \sin \frac{k\pi x}{d},$$

which converges to f in $L_2[0, d]$. In addition, it is not difficult to verify that

$$W_2 = \left\{ \sqrt{\frac{2}{d}} \sin \frac{k\pi x}{d} : k = 1, 2, \ldots \right\} \tag{6.2.14}$$

is an orthonormal family in $L_2[0, d]$ (see Exercise 7). Thus, W_2 is an orthonormal basis of $L_2[0, d]$. We summarize the above discussion in the following theorem.

Theorem 4 $\boxed{\text{Fourier sine series}}$ *Let $d > 0$. Then the family W_2 of sine functions defined by (6.2.14) is an orthonormal basis of $L_2[0, d]$. Consequently, any function $f \in L_2[0, d]$ can be represented by its sine series, namely:*

$$f(x) = (S^s f)(x) = \sum_{k=1}^{\infty} b_k \sin \frac{k\pi x}{d}, \qquad (6.2.15)$$

which converges to $f(x)$ in $L_2[0, d]$, where the sine coefficients b_k are given by

$$b_k = \frac{2}{d} \int_0^d f(x) \sin \frac{k\pi x}{d} dx, k = 1, 2, \ldots. \qquad (6.2.16)$$

Example 4 Let $f(x) = x, 0 \le x \le 1$. Compute the sine series $S^s f$ of f in (6.2.15) with $d = 1$.

Solution For $d = 1$ in (6.2.16),

$$b_k = 2 \int_0^1 f(x) \sin k\pi x \, dx = 2 \int_0^1 x \sin k\pi x \, dx$$

$$= 2 \int_0^1 x \left(\frac{-\cos k\pi x}{k\pi} \right)' dx = \frac{-2}{k\pi} \int_0^1 x (\cos k\pi x)' \, dx$$

$$= \frac{-2}{k\pi} \left\{ \left[x \cos k\pi x \right]_0^1 - \int_0^1 \cos k\pi x \, dx \right\} = \frac{-2}{k\pi} \left\{ \cos k\pi - 0 - \left[\frac{\sin k\pi x}{k\pi} \right]_0^1 \right\}.$$

$$= \frac{-2}{k\pi} \left\{ (-1)^k - \frac{\sin k\pi}{k\pi} + \frac{\sin 0}{k\pi} \right\} = \frac{-2}{k\pi} \{ (-1)^k - 0 + 0 \}$$

$$= \frac{2}{k\pi} (-1)^{k+1} = \frac{2}{\pi} \frac{(-1)^{k+1}}{k}$$

Thus, by (6.2.15), the Fourier sine series of $f(x) = x$ is given by

$$\frac{2}{\pi} \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} \sin k\pi x.$$

∎

In applications, there are other types of cosine and sine series for $f \in L_2[0, d]$ that are more frequently used. Let $g(x)$ be the odd extension of f around $x = d$ (instead of the even extension around 0 as in the derivation of (6.2.10) that leads to (6.2.12)–(6.2.13) of Theorem 3), namely:

$$g(x) = \begin{cases} f(x), & \text{for } 0 \le x < d, \\ -f(2d - x), & \text{for } d < x \le 2d, \\ 0, & \text{for } x = d. \end{cases}$$

Then by Theorem 3, as a function in $L_2[0, 2d]$, $g(x)$ can be expanded as its cosine series, namely:

$$g(x) = (S^c g)(x) = \frac{a_0(g)}{2} + \sum_{k=1}^{\infty} a_k(g) \cos \frac{k\pi x}{2d},$$

where

$$\begin{aligned}
a_k(g) &= \frac{2}{2d} \int_0^{2d} g(x) \cos \frac{k\pi x}{2d} dx \\
&= \frac{1}{d} \int_0^d f(x) \cos \frac{k\pi x}{2d} dx + \frac{1}{d} \int_d^{2d} \left(-f(2d - x) \right) \cos \frac{k\pi x}{2d} dx \\
&= \frac{1}{d} \int_0^d f(x) \cos \frac{k\pi x}{2d} dx + \frac{1}{d} \int_0^d \left(-f(y) \right) \cos \frac{k\pi (2d - y)}{2d} dy \\
&= \frac{1}{d} \int_0^d f(x) \left(1 - (-1)^k \right) \cos \frac{k\pi x}{2d} dx.
\end{aligned}$$

Thus,

$$a_{2k}(g) = 0, \, a_{2k+1}(g) = \frac{2}{d} \int_0^d f(x) \cos \frac{(2k+1)\pi x}{2d} dx, \, k = 0, 1, \ldots.$$

Since the cosine series $S^c g$ of g converges to g in $L_2[0, 2d]$, it converges to f in $L_2[0, d]$. In addition, it is not difficult to verify that

$$W_3 = \left\{ \sqrt{\frac{2}{d}} \cos \frac{(2k+1)\pi x}{2d} : k = 0, 1, \ldots \right\} \tag{6.2.17}$$

is an orthonormal family in $L_2[0, d]$ (see Exercise 9). Hence, W_3 is an orthonormal basis of $L_2[0, d]$. We summarize the above discussion in the following theorem.

Theorem 5 | **Fourier cosine series of Type II** | *Let $d > 0$. The family W_3 defined by (6.2.17) is an orthonormal basis of $L_2[0, d]$. Consequently, any $f \in L_2[0, d]$ can be represented by the cosine series*

$$f(x) = (\widetilde{S}^c f)(x) = \sum_{k=0}^{\infty} a_k(f) \cos \frac{(2k+1)\pi x}{2d}, \tag{6.2.18}$$

which converges to $f(x)$ in $L_2[0, d]$, where the cosine coefficients $a_k(f)$ are given by

$$a_k(f) = \frac{2}{d} \int_0^d f(x) \cos \frac{(2k+1)\pi x}{2d} dx, \, k = 0, 1, \ldots. \tag{6.2.19}$$

Remark 1 It is important to observe that the DCT studied in Sect. 4.3 of Chap. 4 and applied to JPEG image compression in Chap. 5 is the discrete version of the cosine

coefficients $a_k(f)$, $k = 0, 1, \ldots,$ in (6.2.19), and that the inverse DCT is the discrete version of the cosine series $\widetilde{S}^c f$ in (6.2.18). ∎

Example 5 Let $f(x) = x, 0 \le x \le 1$. Compute the cosine series $\widetilde{S}^c f$ of f in (6.2.18) with $d = 1$.

Solution For $d = 1$ in (6.2.19),

$$
\begin{aligned}
a_k(f) &= 2 \int_0^1 f(x) \cos\left(k + \frac{1}{2}\right) \pi x \, dx = 2 \int_0^1 x \cos\left(k + \frac{1}{2}\right) \pi x \, dx \\
&= 2\left[x \frac{\sin\left(k + \frac{1}{2}\right) \pi x}{\left(k + \frac{1}{2}\right) \pi}\right]_0^1 - 2 \int_0^1 \frac{\sin\left(k + \frac{1}{2}\right) \pi x}{\left(k + \frac{1}{2}\right) \pi} \, dx \\
&= \frac{2 \sin\left(k + \frac{1}{2}\right) \pi}{\left(k + \frac{1}{2}\right) \pi} + \frac{2}{\left(\left(k + \frac{1}{2}\right)\pi\right)^2}\left[\cos\left(k + \frac{1}{2}\right) \pi x\right]_0^1 \\
&= \frac{2(-1)^k}{\left(k + \frac{1}{2}\right)\pi} - \frac{2}{\left(\left(k + \frac{1}{2}\right)\pi\right)^2}.
\end{aligned}
$$

Thus, by (6.2.18), the Fourier cosine series of $f(x) = x$ is given by

$$
(\widetilde{S}^c f)(x) = \sum_{k=0}^{\infty} \left(\frac{2(-1)^k}{\left(k + \frac{1}{2}\right) \pi} - \frac{2}{\left(\left(k + \frac{1}{2}\right) \pi\right)^2}\right) \cos\left(k + \frac{1}{2}\right) \pi x.
$$

∎

Finally, let us consider another type of sine series for $f \in L_2[0, d]$. Let $h(x)$ be the even extension of f around $x = d$ (instead of the odd extension around 0 as in the derivation of (6.2.15)–(6.2.16) of Theorem 4), namely:

$$
h(x) = \begin{cases} f(x), & \text{for } 0 \le x \le d, \\ f(2d - x), & \text{for } d < x \le 2d. \end{cases}
$$

Then, by Theorem 4, as a function in $L_2[0, 2d]$, $h(x)$ can be expanded as its sine series, namely:

$$
h(x) = (S^s h)(x) = \sum_{k=1}^{\infty} b_k(h) \sin \frac{k\pi x}{2d},
$$

where, by the symmetry property of $h(x)$,

$$
\begin{aligned}
b_k(h) &= \frac{2}{2d} \int_0^{2d} h(x) \sin \frac{k\pi x}{2d} \, dx \\
&= \frac{1}{d} \int_0^d f(x)\left(1 - (-1)^k\right) \sin \frac{k\pi x}{2d} \, dx \quad\quad (6.2.20)
\end{aligned}
$$

(see Exercise 11). Thus,

$$b_{2k}(h) = 0, \; b_{2k+1}(h) = \frac{2}{d} \int_0^d f(x) \sin \frac{(2k+1)\pi x}{2d} dx, \; k = 0, 1, \ldots.$$

Since the sine series $S^s h$ of h converges to h in $L_2[0, 2d]$, it converges to f in $L_2[0, d]$. In addition, it is not difficult to verify that

$$W_4 = \left\{ \sqrt{\frac{2}{d}} \sin \frac{(2k+1)\pi x}{2d} : k = 0, 1, \ldots \right\} \tag{6.2.21}$$

is an orthonormal family in $L_2[0, d]$ (see Exercise 7). Thus, W_4 is an orthonormal basis of $L_2[0, d]$. We may summarize the above discussion in the following theorem.

Theorem 6 $\boxed{\textbf{Fourier sine series of Type II}}$ *Let $d > 0$. The family W_4 defined by* (6.2.21) *is an orthonormal basis of $L_2[0, d]$. Consequently, any $f \in L_2[0, d]$ can be represented by the sine series*

$$f(x) = (\widetilde{S}^s f)(x) = \sum_{k=0}^{\infty} b_k(f) \sin \frac{(2k+1)\pi x}{2d}, \tag{6.2.22}$$

which converges to $f(x)$ in $L_2[0, d]$, where the sine coefficients $b_k(f)$ are given by

$$b_k(f) = \frac{2}{d} \int_0^d f(x) \sin \frac{(2k+1)\pi x}{2d} dx, \; k = 0, 1, \ldots. \tag{6.2.23}$$

Example 6 Let $f(x) = x, 0 \le x \le 1$. Compute the sine series $\widetilde{S}^s f$ of f in (6.2.22) with $d = 1$.

Solution By (6.2.23) with $d = 1$,

$$
\begin{aligned}
b_k(f) &= 2 \int_0^1 f(x) \sin \left(k + \frac{1}{2} \right) \pi x dx = 2 \int_0^1 x \sin \left(k + \frac{1}{2} \right) \pi x dx \\
&= -2 \left[x \frac{\cos \left(k + \frac{1}{2} \right) \pi x}{\left(k + \frac{1}{2} \right) \pi} \right]_0^1 + 2 \int_0^1 \frac{\cos \left(k + \frac{1}{2} \right) \pi x}{\left(k + \frac{1}{2} \right) \pi} dx \\
&= 0 + \frac{2}{\left(\left(k + \frac{1}{2} \right) \pi \right)^2} \left[\sin \left(k + \frac{1}{2} \right) \pi x \right]_0^1 \\
&= \frac{2(-1)^k}{\left(\left(k + \frac{1}{2} \right) \pi \right)^2}.
\end{aligned}
$$

Thus, by (6.2.22), the Fourier sine series of $f(x) = x$ is given by

$$(\widetilde{S}^s f)(x) = \sum_{k=0}^{\infty} \frac{2(-1)^k}{\left(\left(k + \frac{1}{2}\right) \pi\right)^2} \sin\left(k + \frac{1}{2}\right) \pi x.$$

■

Exercises

Exercise 1 Compute the Fourier cosine and sine series (6.2.1) of each of the following functions:

(a) $g_1(x) = \cos^2 x$,
(b) $g_2(x) = 1 + 2\sin^2 x$,
(c) $g_3(x) = \sin^3 x$,
(d) $g_4(x) = g_1(x) - g_2(x) + g_3(x)$.

Exercise 2 Verify (6.2.5) for any complex numbers $a, b \in \mathbb{C}$.

Exercise 3 Verify that W, defined by (6.2.6), is an orthonormal family of $L_2[-d, d]$.

Exercise 4 Compute the cosine series (6.2.12) with $d = 1$ for each of the following functions:

(a) $f_1(x) = 1, 0 \le x \le 1$,
(b) $f_2(x) = 1, 0 \le x \le \frac{1}{2}$; and $f_2(x) = -1, \frac{1}{2} < x \le 1$,
(c) $f_3(x) = x^2, 0 \le x \le 1$,
(d) $f_4(x) = \sin^2 \pi x, 0 \le x \le 1$.

Exercise 5 Verify that W_1, defined by (6.2.11), is an orthonormal family in $L_2[0, d]$.

Exercise 6 Find the sine series (6.2.15) with $d = 1$ for each of the functions in Exercise 4

Exercise 7 Show that W_2, defined by (6.2.14), is an orthonormal family in $L_2[0, d]$.

Exercise 8 Compute the cosine series (6.2.18) with $d = 1$ for each of the functions in Exercise 4.

Exercise 9 Verify that W_3, defined by (6.2.17), is an orthonormal family in $L_2[0, d]$.

Exercise 10 Compute the sine series (6.2.22) with $d = 1$ for each of the functions in Exercise 4.

Exercise 11 Provide the details for the derivation of (6.2.20).

Exercise 12 Verify that W_4, defined by (6.2.21), is an orthonormal family in $L_2[0, d]$.

Exercise 13 Let $f \in L_2[0, d]$ and $\{a_k(f)\}$ be the sequence defined by (6.2.19). Show that

$$a_k(f) \to 0, \text{ as } k \to \infty.$$

Hint: Apply Parseval's identity with the orthonormal basis W_3, defined by (6.2.17), for $L_2[0, d]$.

Exercise 14 Let $f \in L_2[0, d]$ and $\{b_k(f)\}$ be the sequence defined by (6.2.23). Show that

$$b_k(f) \to 0, \text{ as } k \to \infty.$$

Hint: Apply Parseval's identity with the orthonormal basis W_4, defined by (6.2.21), for $L_2[0, d]$.

6.3 Kernel Methods

For $f \in PC_{2\pi}^\star$, the nth partial sum $(S_n f)(x)$ of the Fourier series expansion (6.1.4) on p. 267 can be formulated as

$$
\begin{aligned}
(S_n f)(x) &= \sum_{k=-n}^{n} c_k e^{ikx} = \sum_{k=-n}^{n} \left(\frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) e^{-ikt} dt \right) e^{ikx} \\
&= \sum_{k=-n}^{n} \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) e^{ik(x-t)} dt \\
&= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) D_n(x-t) dt,
\end{aligned}
\tag{6.3.1}
$$

where

$$
D_n(x) = \sum_{k=-n}^{n} e^{ikx}
\tag{6.3.2}
$$

is called the nth order "**Dirichlet kernel**" .

Definition 1 $\boxed{\text{Convolution}}$ *For the inner-product space $PC_{2\pi}^\star$ with inner product defined in (6.1.1), the operation*

$$
(f * g)(x) = \int_{-\pi}^{\pi} f(t) g(x-t) dt
\tag{6.3.3}
$$

is called the convolution of $f \in PC_{2\pi}^\star$ with $g \in PC_{2\pi}^\star$.

Remark 1 The convolution operator defined by (6.3.3) is commutative, since

$$
\int_{-\pi}^{\pi} f(t) g(x-t) dt = \int_{-\pi}^{\pi} f(x-t) g(t) dt,
$$
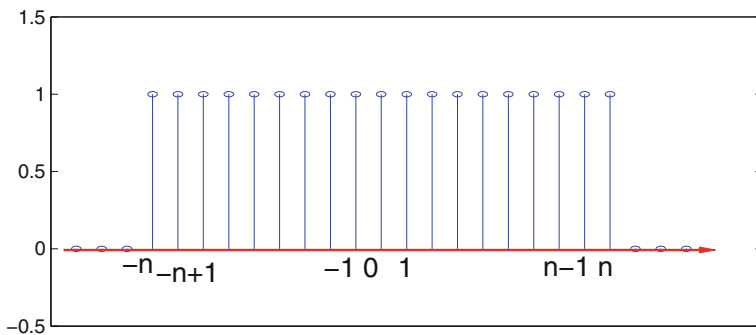
Fig. 6.1 Coefficient plot of $D_n(x)$

which can be easily verified by applying the 2π-periodicity property of $f(x)$ and $g(x)$ in the change of variable $(x - t)$ to t in the integration (where x is fixed) (see Exercise 1). ∎

Remark 2 As shown in (6.3.1), the nth partial sum $(S_n f)(x)$ of the Fourier series of $f \in PC_{2\pi}^\star$ is a $\frac{1}{2\pi}$-multiple of the convolution of f with the nth order Dirichlet kernel, namely, $S_n f = \frac{1}{2\pi} f * D_n$. Hence, if $\frac{1}{2\pi} D_n(x)$ is considered as a (convolution) filter of the signal $f \in PC_{2\pi}^\star$, then it is an "ideal lowpass filter" in that the high-frequency content, $c_k(f)$ for $|k| > n$, is removed, while the low-frequency content, $c_k(f)$ for $k = -n, \ldots, n$, is unaltered. Indeed, the Fourier coefficients c_k of $D_n(x)$ are given by $c_k = 1$ for $|k| \le n$, and $c_k = 0$ for $|k| > n$, as shown in Fig. 6.1. This concept will be elaborated upon in Sect. 7.2. ∎

Theorem 1 *Let f and g be 2π-periodic functions such that $f \in L_2[-\pi, \pi]$ and $g \in L_1[-\pi, \pi]$. Then $f * g \in L_2[-\pi, \pi]$ and*

$$\|f * g\|_2 \le \|f\|_2 \|g\|_1. \tag{6.3.4}$$

Although (6.3.4) is valid for Lebesgue integrable functions as stated in the theorem, we only consider functions in $PC_{2\pi}^*$ in this elementary textbook. Hence, both of the functions $f * g$ and g can be approximated by finite Riemann sums. More precisely,

$$\lim_{n \to \infty} \sum_{k=0}^{n-1} f\left(x - \frac{2\pi k}{n}\right) g\left(\frac{2\pi k}{n}\right) \frac{2\pi}{n} = (f * g)(x); \tag{6.3.5}$$

$$\lim_{n \to \infty} \sum_{k=0}^{n-1} \left| g\left(\frac{2\pi k}{n}\right) \right| \frac{2\pi}{n} = \int_{-\pi}^{\pi} |g(t)| dt = \|g\|_1. \tag{6.3.6}$$

Now, by setting

$$F_{n,k}(x) = f\left(x - \frac{2\pi k}{n}\right) g\left(\frac{2\pi k}{n}\right) \frac{2\pi}{n}$$

in (6.3.5), we have

$$\lim_{n\to\infty} \left\|\sum_{k=0}^{n-1} F_{n,k}\right\|_2 = \left(\lim_{n\to\infty} \int_{-\pi}^{\pi} \left|\sum_{k=0}^{n-1} F_{n,k}(x)\right|^2 dx\right)^{\frac{1}{2}}$$

$$= \left(\int_{-\pi}^{\pi} \lim_{n\to\infty} \left|\sum_{k=0}^{n-1} F_{n,k}(x)\right|^2 dx\right)^{\frac{1}{2}} = \|f * g\|_2.$$

Here, the justification for the validity of interchanging limit and integration is that $\left|\sum_{k=0}^{n-1} F_{n,k}(x)\right|^2$ is nonnegative. It can also be proved rigorously by applying Fatou's lemma, which is one of the most basic theorems in a course of Real Analysis (see Theorem 3 in Sect. 7.6 of Chap. 7).

On the other hand, it follows from the triangle inequality for the L_2-norm, that

$$\left\|\sum_{k=0}^{n-1} F_{n,k}\right\|_2 \leq \sum_{k=0}^{n-1} \|F_{n,k}\|_2 = \sum_{k=0}^{n-1} \left(\int_{-\pi}^{\pi} |F_{n,k}(x)|^2 dx\right)^{\frac{1}{2}}$$

$$= \sum_{k=0}^{n-1} \left\{\left(\int_{-\pi}^{\pi} \left|f\left(x - \frac{2\pi k}{n}\right)\right|^2 dx\right)\right.$$

$$\left.\left(\left|g(\frac{2\pi k}{n})\frac{2\pi}{n}\right|^2\right)\right\}^{\frac{1}{2}}$$

$$= \sum_{k=0}^{n-1} \left(\int_{-\pi}^{\pi} |f(x)|^2 dx\right)^{\frac{1}{2}} \left|g\left(\frac{2\pi k}{n}\right)\right| \frac{2\pi}{n}$$

$$= \|f\|_2 \sum_{k=0}^{n-1} \left|g\left(\frac{2\pi k}{n}\right)\right| \frac{2\pi}{n} \to \|f\|_2 \|g\|_1,$$

for $n \to \infty$, by (6.3.6). Here, we have applied the 2π-periodic property of $f(x)$ to obtain

$$\int_{-\pi}^{\pi} \left|f\left(x - \frac{2\pi k}{n}\right)\right|^2 dx = \int_{-\pi}^{\pi} |f(x)|^2 dx.$$

This completes the proof of (6.3.4).   ∎

We remark that Theorem 1 can also be proved by applying the more general Minkowski inequality for integrals to be discussed in the Appendix of Chap. 7. In the following, we derive an explicit formulation of the Dirichlet kernel.

Theorem 2 *For $n = 1, 2, 3, \ldots$, the Dirichlet kernel $D_n(x)$, as defined in (6.3.2), is given by*

$$D_n(x) = \frac{\sin(n + \frac{1}{2})x}{\sin \frac{x}{2}}. \tag{6.3.7}$$

The formula (6.3.7) can be derived as follows:

$$
\begin{aligned}
D_n(x) &= e^{-inx} \sum_{k=0}^{2n} e^{ikx} = e^{-inx} \frac{1 - e^{i(2n+1)x}}{1 - e^{ix}} \\
&= \frac{e^{-inx} - e^{i(n+1)x}}{e^{i\frac{x}{2}}(e^{-i\frac{x}{2}} - e^{i\frac{x}{2}})} = \frac{e^{-i(n+\frac{1}{2})x} - e^{i(n+\frac{1}{2})x}}{e^{-i\frac{x}{2}} - e^{i\frac{x}{2}}} \\
&= \frac{(-2i)\sin\left(n + \frac{1}{2}\right)x}{(-2i)\sin \frac{x}{2}} = \frac{\sin\left(n + \frac{1}{2}\right)x}{\sin \frac{x}{2}}. \qquad \blacksquare
\end{aligned}
$$

Remark 3 From the definition in (6.3.2), it is clear that $D_n(0) = 2n + 1$. Hence, the formula in (6.3.7) of $D_n(x)$ for $x \neq 0$ also applies to $x = 0$ by applying L'Hospital's rule, namely:

$$
\begin{aligned}
\lim_{x \to 0} \frac{\sin(n + \frac{1}{2})x}{\sin \frac{x}{2}} &= \lim_{x \to 0} \frac{(n + \frac{1}{2})\cos(n + \frac{1}{2})x}{\frac{1}{2}\cos \frac{x}{2}} \\
&= (2n + 1)\frac{\cos 0}{\cos 0} = 2n + 1.
\end{aligned}
$$

Examples of graphs of $D_n(x)$ are plotted in Fig. 6.2. $\blacksquare$

Next, we introduce the notion of **Césaro means** (that is, averages) of Dirichlet's kernels $D_j(x)$ to formulate the nth order Fejér's kernel, namely:
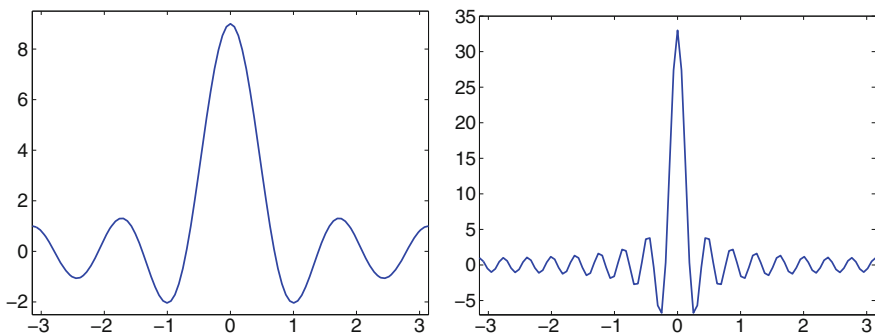
Fig. 6.2 Dirichlet's kernels $D_n(x)$ for $n = 4$ (*on left*) and $n = 16$ (*on right*)
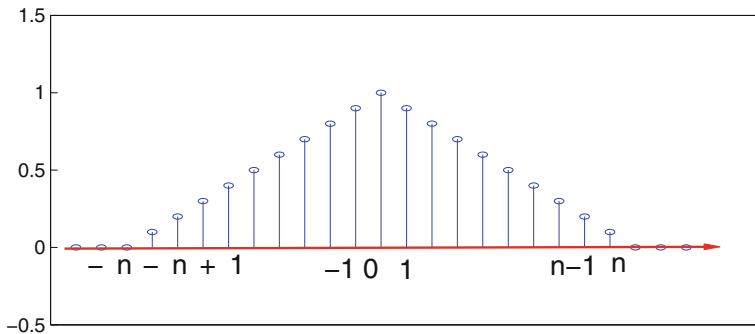
Fig. 6.3 Coefficient plot of $\sigma_n(x)$

$$\sigma_n(x) = \frac{1}{n+1} \sum_{j=0}^{n} D_j(x), \tag{6.3.8}$$

which is another lowpass filter with coefficient plot shown in Fig. 6.3. For convenience, we set $z = e^{ix}$ and observe that

$$D_j(x) = \frac{1 - z^{2j+1}}{z^j(1-z)} = \frac{1}{z-1}\left(z^{j+1} - \frac{1}{z^j}\right).$$

Thus, we have

$$
\begin{aligned}
(n+1)\sigma_n(x) &= D_0(x) + D_1(x) + \cdots + D_n(x)\\
&= \frac{1}{z-1}(z-1) + \frac{1}{z-1}\left(z^2 - \frac{1}{z}\right) + \cdots + \frac{1}{z-1}\left(z^{n+1} - \frac{1}{z^n}\right)\\
&= \frac{1}{z-1}\left(z + z^2 + \cdots + z^{n+1} - 1 - \frac{1}{z} - \cdots - \frac{1}{z^n}\right)\\
&= \frac{1}{z-1}\left(z - \frac{1}{z^n}\right)\left(1 + z + z^2 + \cdots + z^n\right)\\
&= \frac{1}{(z-1)^2}\left(z - \frac{1}{z^n}\right)(z^{n+1} - 1)\\
&= \frac{(z^{n+1} - 1)^2}{z^n(z-1)^2} = \frac{\left(z^{\frac{n+1}{2}} - z^{-\frac{n+1}{2}}\right)^2}{\left(z^{\frac{1}{2}} - z^{-\frac{1}{2}}\right)^2}\\
&= \left(\frac{\sin\frac{(n+1)x}{2}}{\sin\frac{x}{2}}\right)^2.
\end{aligned}
$$

In the following, we summarize the important properties of the Fejér kernel.

Theorem 3 | **Fejér's kernel** | *For $n = 1, 2, \ldots$, the nth order Fejér kernel $\sigma_n(x)$, as defined in* (6.3.8)*, can be formulated as*

$$\sigma_n(x) = \frac{1}{n+1} \left(\frac{\sin \frac{(n+1)x}{2}}{\sin \frac{x}{2}} \right)^2. \tag{6.3.9}$$

Furthermore, the sequence $\{\sigma_n(x)\}$ constitutes a positive "approximate identity", meaning that it has the following properties:

(a) $\sigma_n(x) \geq 0$, *for all x*;

(b) $\dfrac{1}{2\pi} \displaystyle\int_{-\pi}^{\pi} \sigma_n(x)dx = 1$;

(c) *For any positive number δ with $0 < \delta < \pi$,*

$$\sup_{\delta \leq |x| \leq \pi} \sigma_n(x) \to 0, \text{ as } n \to \infty; \tag{6.3.10}$$

(d) *For any $f \in L_2[-\pi, \pi]$ with 2π-periodic extension also denoted by f,*

$$\frac{1}{2\pi} \| f * \sigma_n \|_2 \leq \| f \|_2.$$

The first property is evident from (6.3.9), and the second property follows from the definitions of $D_k(x)$ and $\sigma_n(x)$, while the third property follows from the fact that $|\sin y| \leq 1$ for all $y \in \mathbb{R}$ and $|\sin \frac{x}{2}| \geq |\sin \frac{\delta}{2}| > 0$ for $\delta \leq |x| \leq \pi$ (see Exercise 2). The fourth property is a consequence of Theorem 1 together with properties (a) and (b). Examples of graphs of $\sigma_n(x)$ are illustrated in Fig. 6.4. ∎

In view of the convolution formulation (in terms of the Dirichlet kernels and f) for the partial sum $(S_n f)(x)$ of the Fourier series $(Sf)(x)$ in (6.3.1), we consider the Césaro means of $(S_n f)(x)$ as convolution with the Fejér kernels $\sigma_n(x)$, as follows.
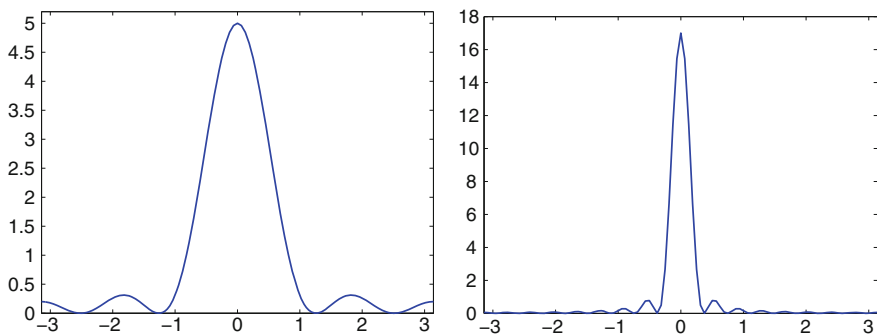
Fig. 6.4 Fejer's kernels $\sigma_n(x)$ for $n = 4$ (*on left*) and $n = 16$ (*on right*)

Definition 2 | **Césaro means** | *The nth-order Césaro means* $(C_n f)(x)$ *of* $f \in PC_{2\pi}^{\star}$
is defined by

$$(C_n f)(x) = \frac{(S_0 f)(x) + \cdots + (S_n f)(x)}{n + 1}. \tag{6.3.11}$$

Observe that $\mathbb{V}_{2n+1}$, defined by (6.1.2) on p.267, is a subspace of $PC_{2\pi}^{\star}$. Since $(C_n f)(x)$ is a linear combination of $(S_j f)(x)$, $0 \leq j \leq n$, and since each $(S_j f)(x)$ is in $\mathbb{V}_{2n+1}$, the Césaro means $(C_n f)(x)$ is also in $\mathbb{V}_{2n+1}$.

Remark 4 It follows from (6.3.1) and (6.3.8) that

$$(C_n f)(x) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t)\sigma_n(x - t)dt = \frac{1}{2\pi}(f * \sigma_n)(x), \tag{6.3.12}$$

where $\sigma_n(x)$ is the nth order Fejér kernel. Hence, by applying property (d) of Theorem 3, we also have, for any $g \in PC_{2\pi}^{\star}$,

$$\|C_n g\|_2 = \frac{1}{2\pi}\|g * \sigma_n\|_2 \leq \|g\|_2. \tag{6.3.13}$$

∎

Exercises

Exercise 1 Show that the convolution operation defined in (6.3.3) for $PC_{2\pi}^{\star}$ is commutative.

Exercise 2 Show that, for $0 < \delta \leq \pi$,

$$\left| \sin \frac{x}{2} \right| \geq \left| \sin \frac{\delta}{2} \right| \geq \frac{\delta}{\pi}$$

for all x, with $\delta \leq x \leq \pi$.

Exercise 3 Let $D_n(t)$ be the Dirichlet kernel. Show that $\int_0^{\pi} D_n(t)dt = \pi$.

Exercise 4 Let $(C_n f)(x)$ be the nth order Césaro means of the partial sums $S_0 f, \ldots, S_n f$ of the Fourier series $(Sf)(x)$, as defined in (6.3.11). Derive the convolution formula $(C_n f)(x) = \frac{1}{2\pi}(f * \sigma_n)(x)$ in (6.3.12), where $\sigma_n(x)$ is the nth order Fejér kernel.

Exercise 5 From the formula (6.3.9) of the Fejér kernel, compute the value of $\sigma_n(x)$ at $x = 0$.

Exercise 6 Does the sequence of Césaro means of the divergent sequence $\{x_n\}$, $x_n = (-1)^n$, converge? If so, compute its limit.

Exercise 7 As a continuation of Exercise 6, let $s_n = \sum_{k=0}^{n} x_k$. Show that $\{s_n\}$ is divergent although the sequence of its Césaro means converges to $\frac{1}{2}$.

Exercise 8 (For advanced students) If a sequence $\{x_n\}$ of complex numbers converges to some limit $c \in \mathbb{C}$ as $n \to \infty$, show that the sequence of its Césaro means

$$y_n = \frac{x_0 + \cdots + x_n}{n+1}$$

also converges to c.

6.4 Convergence of Fourier Series

As mentioned previously, when a function $f \in L_2[-\pi, \pi]$ is considered as an analog signal, its Fourier coefficients $c_k = c_k(f)$, $k \in \mathbb{Z}$, reveal its frequency content, with the coefficient c_k representing the content at $(k/2\pi)$Hz. Therefore, the separation of $f(x)$ into frequency components $c_k e^{ikx}$, $a_k \cos kx$, or $b_k \sin kx$ facilitates the application of digital signal processing (DSP) of $f(x)$. However, it is important that the (modified) function $f(x)$ can be recovered from the (processed) sequence $\{c_k\}$, or equivalently, from the Fourier series $(Sf)(x)$. The first theorem in this section assures that for $f \in PC_{2\pi}^{\star}$, $(Sf)(x_0)$ converges to $f(x_0)$, under the assumption that $f(x)$ has both one-sided derivatives at $x = x_0$.

Theorem 1 $\boxed{\text{Pointwise convergence}}$ *Let $f(x) \in PC[-\pi, \pi]$ and $x_0 \in [-\pi, \pi]$, and suppose that both of the one-sided derivatives*

$$f'(x_0^+) = \lim_{h \to 0^+} \frac{f(x_0 + h) - f(x_0^+)}{h}, \; f'(x_0^-) = \lim_{h \to 0^+} \frac{f(x_0 - h) - f(x_0^-)}{-h},$$

of $f(x)$ at $x_0 \in [-\pi, \pi]$ exist. Then the Fourier series of f converges to $\frac{1}{2}(f(x_0^-) + f(x_0^+))$ at $x = x_0$.

In the above theorem, $f(x_0^+)$ and $f(x_0^-)$ denote the right-hand and left-hand limits of f at x_0. Moreover, in view of the 2π-periodic extension of $f(x)$, we have

$$\begin{cases} f(\pi^+) = f(-\pi^+), & f(-\pi^-) = f(\pi^-); \\ f'(\pi^+) = f'(-\pi^+), & f'(-\pi^-) = f'(\pi^-). \end{cases} \tag{6.4.1}$$

$\boxed{\text{Proof of Theorem1}}$ Consider the 2π-periodic extension of $f \in PC[-\pi, \pi]$ to $f \in PC_{2\pi}^{\star}$. Then by (6.3.1) on p.288, we have

$$(S_n f)(x) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x - t) D_n(t) dt$$

$$= \frac{1}{2\pi} \left(\int_{-\pi}^{0} + \int_{0}^{\pi} \right) \left(f(x - t) D_n(t) \right) dt$$

$$= \frac{1}{2\pi} \int_{0}^{\pi} \left(f(x + t) + f(x - t) \right) D_n(t) dt,$$

where the change of the variable of integration $u = -t$ for the integral over $[-\pi, 0]$ and the fact that $D_n(-t) = D_n(t)$ have been applied. Notice that

$$\frac{1}{2\pi} \int_{0}^{\pi} D_n(t) dt = \frac{1}{2}$$

(see Exercise 3 of previous section). Thus, for any $x = x_0$, we have

$$(S_n f)(x_0) - \frac{1}{2} \left(f(x_0^-) + f(x_0^+) \right)$$

$$= \frac{1}{2\pi} \int_{0}^{\pi} \left(f(x_0 + t) + f(x_0 - t) - \left(f(x_0^-) + f(x_0^+) \right) \right) D_n(t) dt$$

$$= \int_{0}^{\pi} h(t) \sin \left(n + \frac{1}{2} \right) t \, dt,$$

where

$$h(t) = \frac{1}{2\pi \sin \frac{t}{2}} \left(f(x_0 + t) - f(x_0^+) + f(x_0 - t) - f(x_0^-) \right).$$

By the assumption that $f'(x_0^+)$ and $f'(x_0^-)$ exist, and the fact that $\frac{t}{\pi} \leq \sin \frac{t}{2}$ for all $0 \leq t \leq \pi$ (see Exercise 2), we may conclude that $h(t)$ is bounded on $[0, \pi]$ (see Exercise 3). Thus, $h(t) \in L_2[0, \pi]$. Hence,

$$\lim_{n \to \infty} (S_n f)(x_0) - \frac{1}{2} \left(f(x_0^-) + f(x_0^+) \right) = \lim_{n \to \infty} \int_{0}^{\pi} h(t) \sin(n + \frac{1}{2}) t \, dt = 0,$$

where the last equality follows from Parseval's identity for $h(t)$ with the orthonormal basis $\left\{ \sqrt{\frac{2}{\pi}} \sin(k + \frac{1}{2}) x : k = 0, 1, \ldots \right\}$ of $L_2[0, \pi]$ (see Exercise 14 of Sect. 6.2). ∎

Since $h(t)$ is also in $L_1[0, \pi]$, the last equality in the above proof also follows from the Riemann-Lebesgue lemma, which states that

$$\lim_{\omega \to \infty} \int_{0}^{\pi} g(t) e^{-i\omega t} dt = 0 \tag{6.4.2}$$

for any $g(t) \in L_1[0, \pi]$ (see Exercise 4 for the proof of (6.4.2)).

If $f(x)$ is continuous at $x \in (-\pi, \pi)$, then $\frac{1}{2} \left(f(x^-) + f(x^+) \right) = f(x)$; and if $f(x)$ is continuous at π and $-\pi$ and $f(-\pi) = f(\pi)$, then

$$\frac{1}{2}\left(f(\pi^-) + f(\pi^+)\right) = f(\pi), \frac{1}{2}\left(f(-\pi^-) + f(-\pi^+)\right) = f(-\pi).$$

Furthermore, if $f(x)$ is differentiable at x, then $f'(x^+)$ and $f'(x^-)$ exist. Therefore, for any $f \in PC_{2\pi}^*$, the Fourier series

$$(Sf)(x) = \sum_{k=-\infty}^{\infty} c_k(f)e^{ikx}$$

or

$$(Sf)(x) = \frac{a_0(f)}{2} + \sum_{k=1}^{\infty} a_k(f)\cos kx + b_k(f)\sin kx$$

converges to $f(x)$ at any x, where $f(x)$ is continuous, provided that both of the one-sided derivatives $f'(x^+)$ and $f'(x^-)$ exist.

Example 1 Let $f(x) = |x|$, $-\pi \le x \le \pi$. Discuss the pointwise convergence of the Fourier series of f.

Solution It is clear that the function $f(x) = |x|$ is continuous on $[-\pi, \pi]$ and differentiable on $(-\pi, 0)$ and $(0, \pi)$, with

$$f'(x) = 1 \text{ for } 0 < x < \pi;$$
$$f'(x) = -1 \text{ for } -\pi < x < 0.$$

Furthermore, the one-sided derivatives of $f(x)$ also exist at $x = -\pi, 0, \pi$, with

$$f'(0^+) = 1, f'(0^-) = -1;$$
$$f'(\pi^-) = 1, f'(-\pi^+) = -1,$$

while, by the second line in (6.4.1),

$$f'(-\pi^-) = f'(\pi^-) = 1, f'(\pi^+) = f'(-\pi^+) = -1.$$

Therefore, by Theorem 1, the Fourier series of $f(x) = |x|$ converges to $f(x)$ for all $x \in [-\pi, \pi]$. Here, we have used the fact that

$$\frac{1}{2}\left(f(-\pi^-) + f(-\pi^+)\right) = f(-\pi);$$
$$\frac{1}{2}\left(f(\pi^-) + f(\pi^+)\right) = f(\pi),$$

since the 2π-periodic extension of $f(x) = |x|$ is continuous at the two end-points $x = -\pi, \pi$. See Fig. 6.5 for the partial sums of the Fourier series. ∎
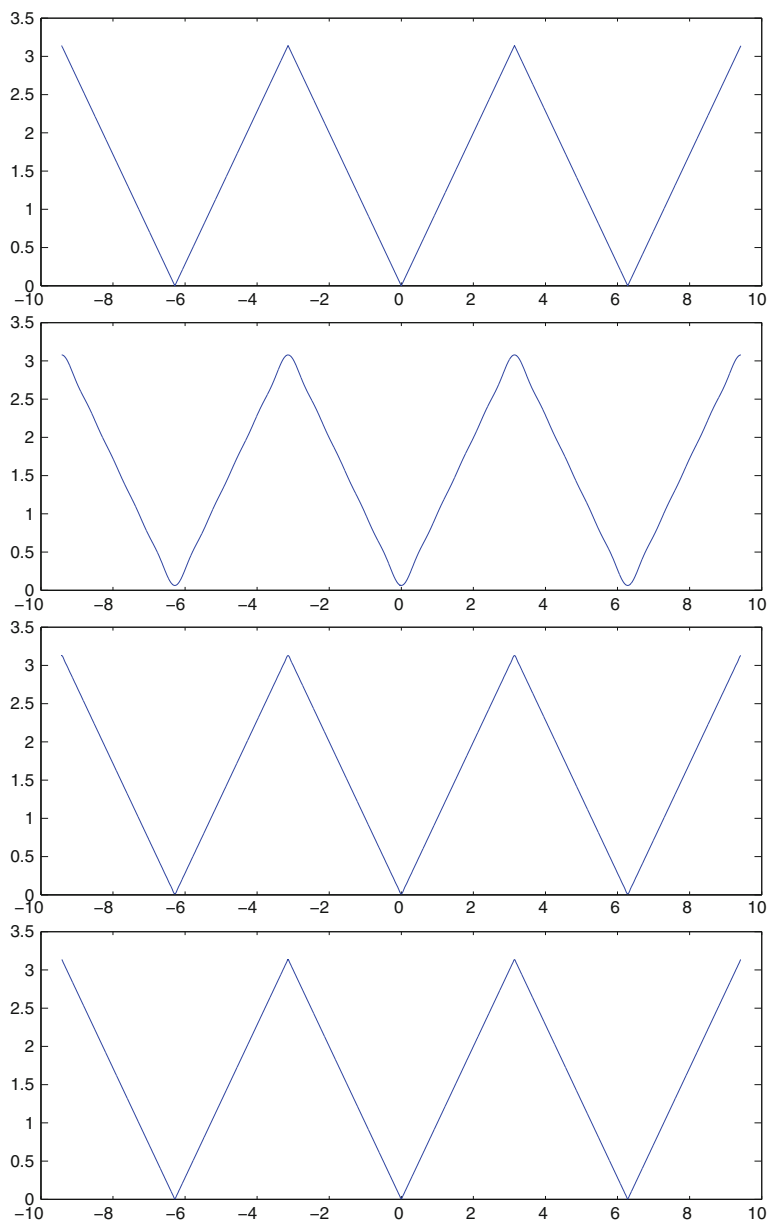
Fig. 6.5 From top to bottom: $f(x) = |x|$ and its 2π-extension, $S_{10}f$, $S_{50}f$ and $S_{100}f$

Example 2 Let $f(x) = \chi_{[-\pi/2,\pi/2]}(x)$, $-\pi \le x \le \pi$ be the function $g_1(x)$ considered in Example 2 on p.269; that is, $f(x) = 1$ for $|x| \le \dfrac{\pi}{2}$ and $f(x) = 0$ for $\dfrac{\pi}{2} < |x| \le \pi$. Discuss the pointwise convergence of the Fourier series of f.

Solution The function $f(x)$ is clearly in $PC[-\pi, \pi]$ and differentiable, with $f'(x) = 0$ for all $x \in [-\pi, \pi]$ except $x = -\dfrac{\pi}{2}, \dfrac{\pi}{2}$, and with

$$f'\big((-\pi/2)^-\big) = f'\big((-\pi/2)^+\big) = 0,\ f'\big((\pi/2)^-\big) = f'\big((\pi/2)^+\big) = 0$$

as well. In addition, the 2π-periodic extension is continuous at the two end-points $x = -\pi, \pi$. Therefore, it follows from Theorem 1 that the Fourier series $(Sf)(x)$ of $f(x)$ converges to $f(x)$ at each $x \in [-\pi, \pi]$ with $x \neq \pm\dfrac{\pi}{2}$, and to

$$\frac{1}{2}\Big(f\big((\pi/2)^-\big) + f\big((\pi/2)^+\big)\Big) = \frac{1}{2}(1 + 0) = \frac{1}{2},$$
$$\frac{1}{2}\Big(f\big((-\pi/2)^-\big) + f\big((-\pi/2)^+\big)\Big) = \frac{1}{2}(0 + 1) = \frac{1}{2}$$

at $x = \dfrac{\pi}{2}$ and $x = -\dfrac{\pi}{2}$, respectively. See Fig. 6.6 for the partial sums of the Fourier series. ∎

Observe that in the graphs of $(S_n f)(x)$ for $n = 10, 50, 100$, as plotted in Fig. 6.6, there are ripples near the points of jump discontinuities of $f(x)$ at $x = -\frac{\pi}{2}$ and $x = \frac{\pi}{2}$. This is due to the **Gibbs phenomenon**, which is explained as follows.

Recall from Example 2 on p.269 that the Fourier coefficients $c_k = c_k(f)$ are given by

$$c_0 = \frac{1}{2},\ c_{2\ell} = 0,\ c_{2\ell-1} = \frac{(-1)^{\ell-1}}{(2\ell - 1)\pi},\ \ell = 1, 2, \ldots.$$

Thus, the $(2n - 1)$th- and $(2n)$th-order partial sums of the Fourier series $(Sf)(x)$ of f are the same, namely:

$$(S_{2n} f)(x) = (S_{2n-1} f)(x) = \frac{1}{2} + \frac{2}{\pi} \sum_{\ell=1}^{n} (-1)^{\ell-1} \frac{\cos(2\ell - 1)x}{2\ell - 1},$$

so that

$$(S_{2n} f)\left(\frac{\pi}{2} - \frac{\pi}{2n}\right) = \frac{1}{2} + \frac{2}{\pi} \sum_{\ell=1}^{n} \frac{\sin \frac{2\ell-1}{2n}\pi}{2\ell - 1}$$

$$= \frac{1}{2} + \frac{1}{\pi} \frac{\pi}{n} \sum_{\ell=1}^{n} \frac{\sin \frac{2\ell-1}{2n}\pi}{\frac{2\ell-1}{2n}\pi}.$$

On the other hand, observe that

$$\frac{\pi}{n} \sum_{\ell=1}^{n} \frac{\sin \frac{2\ell-1}{2n}\pi}{\frac{2\ell-1}{2n}\pi}$$
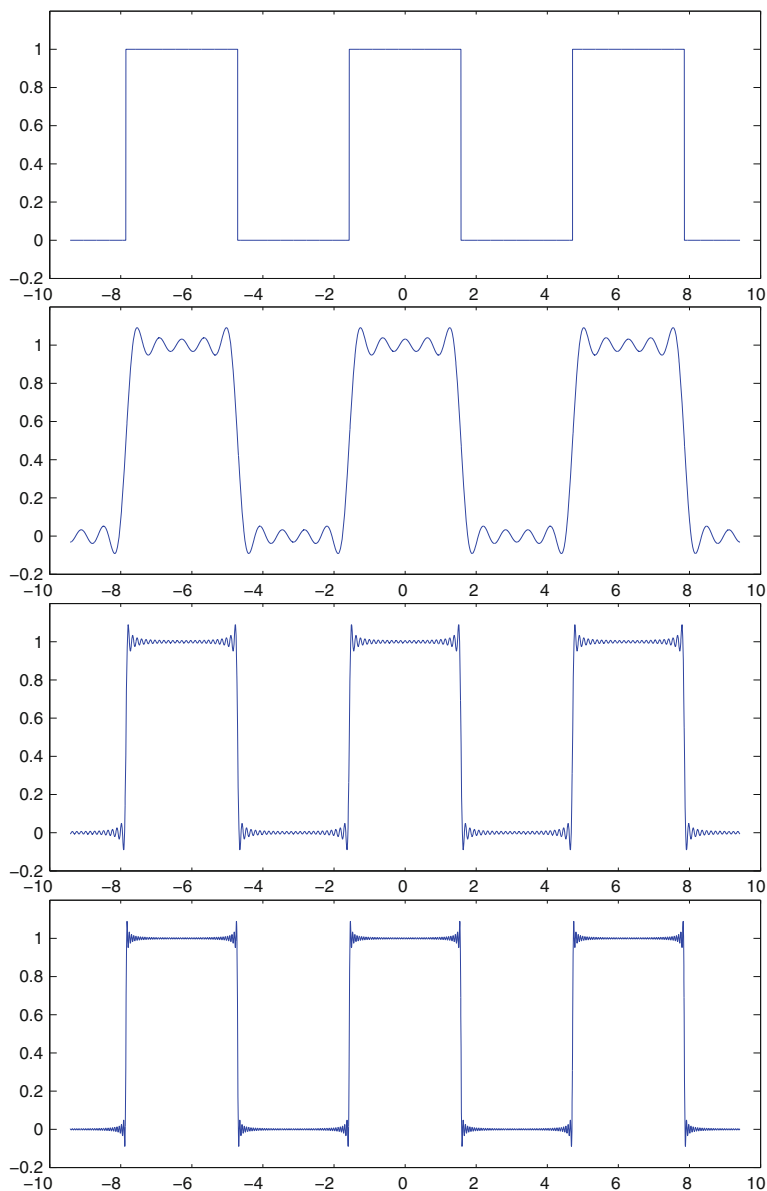
Fig. 6.6 From top to bottom: $f(x) = \chi_{[-\pi/2,\pi/2]}(x)$, $-\pi \le x \le \pi$ and its 2π-extension, $S_{10} f$, $S_{50} f$ and $S_{100} f$

is a Riemann sum of the integral $\int_0^\pi \dfrac{\sin x}{x} dx$. Thus, we have

$$\lim_{n\to\infty} (S_{2n} f) \left(\frac{\pi}{2} - \frac{\pi}{2n}\right) = \frac{1}{2} + \frac{1}{\pi} \int_0^\pi \frac{\sin x}{x} dx.$$

Similarly, it can be shown that

$$\lim_{n\to\infty} (S_{2n-1} f) \left(\frac{\pi}{2} - \frac{\pi}{2n-1}\right) = \frac{1}{2} + \frac{1}{\pi} \int_0^\pi \frac{\sin x}{x} dx.$$

Therefore, for large values of n, we have

$$(S_n f) \left(\frac{\pi}{2} - \frac{\pi}{n}\right) \approx \frac{1}{2} + \frac{1}{\pi} \int_0^\pi \frac{\sin x}{x} dx = \frac{1}{2} + \frac{1}{2} + 0.089490\cdots \approx 1 + 0.09.$$

Notice that $f\left(\left(\frac{\pi}{2}\right)^-\right) = 1$ and the discontinuity jump of f at $\frac{\pi}{2}$ is 1. Therefore, for large n, we have

$$(S_n f) \left(\frac{\pi}{2} - \frac{\pi}{n}\right) - f\left(\left(\frac{\pi}{2}\right)^-\right) \approx 1 \times 0.09.$$

This means that the graph of $(S_n f)(x)$ has an overshoot of about 9% of the jump $f(x_0^+) - f(x_0^-) = 1$ near the point $x_0 = \frac{\pi}{2}$ of jump discontinuity of $f(x)$ (see Fig. 6.7). On the other hand, there is an undershoot of the same amount at $x_0 = \frac{\pi}{2}$ in the other direction. These overshoot and undershoot never disappear as the order of the partial sums $S_n f$ increases. This is called the Gibbs phenomenon.

In general, if f is discontinuous at any point $x_0 \in [-\pi, \pi]$ with the discontinuity jump

$$J = f(x_0^+) - f(x_0^-),$$
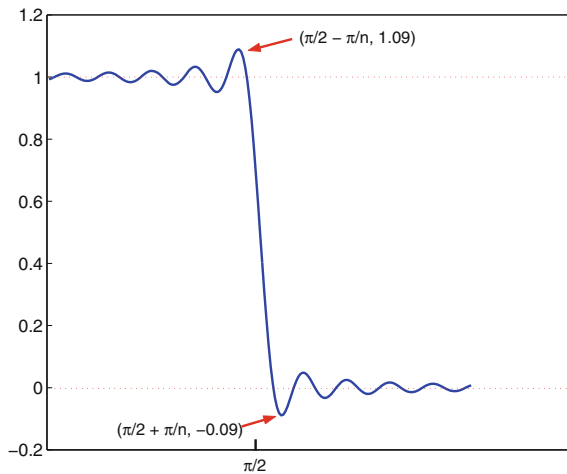
Fig. 6.7 Gibbs phenomenon

then

$$\lim_{n\to\infty} (S_n f)\left(x_0 + \frac{\pi}{n}\right) - f(x_0^+) = J \times 9\,\%,$$

$$\lim_{n\to\infty} (S_n f)\left(x_0 - \frac{\pi}{n}\right) - f(x_0^-) = -J \times 9\,\%.$$

∎

We now turn to the study of convergence of Fourier series in the $\|\cdot\|_2$-norm, induced by the inner product $\langle\cdot,\cdot\rangle$, via the density of trigonometric polynomials (that is, the union of the subspaces $\mathbb{V}_{2n+1}$ defined by (6.1.2) on p.267 for $n = 1, 2, \ldots$) in $PC_{2\pi}^\star$. In the following, we will prove the density result for a stronger norm than the inner product space norm $\|\cdot\|_2$, namely the supremum norm (also called uniform norm), defined by

$$\|f\|_\infty = \sup_x |f(x)| \tag{6.4.3}$$

for $f \in \mathbb{C}_{2\pi}^\star$, where $\mathbb{C}_{2\pi}^\star$ denotes the normed linear space of 2π-periodic continuous functions with norm $\|\cdot\|_\infty$ defined in (6.4.3) (see (1.2.16) on p.20 for the more general notion of *ess sup* norm).

To derive the density result mentioned above, we use the properties of Fejér's kernel as stated in Theorem 3 on p.293 of the previous section.

Theorem 2 $\boxed{\text{Convergence of Céesaro means}}$ *For any $f \in PC_{2\pi}^\star$, its nth Césaro means $C_n f$ is a trigonometric polynomial in $\mathbb{V}_{2n+1}$, and*

$$\|f - C_n f\|_2 \to 0, \text{ as } n \to \infty. \tag{6.4.4}$$

Moreover, if $f \in C_{2\pi}^\star$, then

$$\|f - C_n f\|_\infty \to 0, \text{ as } n \to \infty. \tag{6.4.5}$$

Proof Since the proof of (6.4.4) will depend on (6.4.5), we first prove (6.4.5). Let $\epsilon > 0$ be arbitrarily given. Then for $f \in C_{2\pi}^\star$, since $[-\pi, \pi]$ is a compact interval, $f(x)$ is bounded and uniformly continuous on $[-\pi, \pi]$; that is, $M = \|f\|_\infty < \infty$, and there is a $\delta > 0$ such that

$$|f(x) - f(t)| < \frac{\epsilon}{2} \tag{6.4.6}$$

for all x, t with $|x - t| < \delta$. By the third property of the Fejér kernels $\sigma_n(x)$ in Theorem 3 on p.293, there exists an integer N such that

$$\sup_{\delta \le |x-t| \le \pi} \sigma_n(x - t) \le \frac{\epsilon}{4M} \tag{6.4.7}$$

for all $n \geq N$. In addition, by applying properties (b) and (a) of $\sigma_n(x)$ in the same theorem consecutively, we have

$$\left| f(x) - (C_n f)(x) \right| = \left| \frac{1}{2\pi} \int_{-\pi}^{\pi} (f(x) - f(t))\sigma_n(x - t)dt \right|$$

$$\leq \frac{1}{2\pi} \int_{-\pi}^{\pi} |f(x) - f(t)|\sigma_n(x - t)dt.$$

Hence, partitioning the integral over $[-\pi, \pi]$ into two integrals over $0 \leq |t - x| < \delta$ and $\delta \leq |t - x| \leq \pi$, and applying (6.4.6) and (6.4.7), we have, for all $n \geq N$ and all $x \in [-\pi, \pi]$,

$$|f(x) - (C_n f)(x)| \leq \frac{1}{2\pi} \int_{|x-t|<\delta} |f(x) - f(t)|\sigma_n(x - t)dt$$

$$+ \frac{1}{2\pi} \int_{|x-t|\geq\delta} |f(x) - f(t)|\sigma_n(x - t)dt$$

$$< \frac{\epsilon}{2} \left(\frac{1}{2\pi} \int_{-\pi}^{\pi} \sigma_n(x - t)dt \right) + \frac{1}{2\pi} \int_{-\pi}^{\pi} 2M \cdot \frac{\epsilon}{4M}dt$$

$$= \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

That is, $\| f - C_n f \|_\infty < \epsilon$ for all $n \geq N$, or equivalently,

$$\| f - C_n f \|_\infty \to 0, \text{ as } n \to \infty.$$

This completes the proof of (6.4.5).

In general, if $f \in PC_{2\pi}^\star$, but f is not necessarily continuous, it can be proved that there exists a continuous function $\tilde{f} \in C_{2\pi}^\star$ such that

$$\| f - \tilde{f} \|_2 < \epsilon$$

(see Exercise 1). Hence,

$$\| f - C_n f \|_2 \leq \| f - \tilde{f} \|_2 + \| \tilde{f} - C_n \tilde{f} \|_2 + \| C_n \tilde{f} - C_n f \|_2$$

$$\leq \epsilon + \sqrt{2\pi} \| \tilde{f} - C_n \tilde{f} \|_\infty + \| \tilde{f} - f \|_2$$

$$\leq 2\epsilon + \sqrt{2\pi} \| \tilde{f} - C_n \tilde{f} \|_\infty,$$

where the second inequality follows from (6.3.13) on p.294 with $g = \tilde{f} - f$, and the fact that

$$\| h \|_2 = \left(\int_{-\pi}^{\pi} |h(x)|^2 dx \right)^{\frac{1}{2}} \leq \sqrt{2\pi} \| h \|_\infty$$

for any $h \in PC_{2\pi}^\star$. Since $\widetilde{f} \in C_{2\pi}^\star$, the proof of (6.4.4) is complete by applying (6.4.5). ∎

Remark 1 | **Fourier series does converge in the L_2 − norm** | For $f \in PC_{2\pi}^\star$, since $(C_n f)(x) \in \mathbb{V}_{2n+1}$ and $(S_n f)(x)$ is the best L_2-approximation to f from $\mathbb{V}_{2n+1}$ (see (6.1.8) on p.270), we have

$$\| f - S_n f \|_2 \le \| f - C_n f \|_2 \to 0, \text{ as } n \to \infty.$$

This validates (6.1.9) on p.270 and completes the proof of Theorem 1 in Sect. 6.1. ∎

<div align="center">

Exercises

</div>

Exercise 1 Let $f \in PC[a, b]$ and $\epsilon > 0$. Show that there exists an $\widetilde{f} \in C[a, b]$ such that $\int_a^b |f(x) - \widetilde{f}(x)|^2 dx < \epsilon$.

Exercise 2 Show that $|t| \le \frac{\pi}{2} |\sin t|$ for $|t| \le \frac{\pi}{2}$.

Exercise 3 Show that $h(t)$ in the proof of Theorem 1 is bounded on $[0, \pi]$ and is in $PC[0, \pi]$.

Exercise 4 Show that (6.4.2) holds for any piecewise constant function $g(x) = \sum_{k=1}^{n} c_k \chi_{[a_{k-1}, a_k)}(x)$ on $[0, \pi]$, where $c_k \in \mathbb{R}$, $0 = a_0 < a_1 < \cdots < a_n = \pi$. Then conclude that (6.4.2) holds for any $g \in L_1[0, \pi]$ by the fact that g can be approximated arbitrarily well in $L_1[0, \pi]$ by piecewise constant functions.

Exercise 5 Let $f(x) = x^2, -\pi \le x \le \pi$. Discuss the pointwise convergence of the Fourier series of f by following the solutions of Examples 1–2.

Exercise 6 Let $g(x) = \pi + x$ for $-\pi \le x \le 0$ and $g(x) = \pi - x$ for $0 < x \le \pi$. Discuss the pointwise convergence of the Fourier series of g by following the solutions of Examples 1–2.

Exercise 7 Let $f(x)$ be the function defined on $[-\pi, \pi]$ with $f(x) = -1$ for $-\pi \le x \le 0$ and $f(x) = 1$ for $0 < x \le \pi$. Use Matlab to draw the graphs of partial sums $(S_n f)$ of the Fourier series of f and observe the Gibbs phenomenon.

Exercise 8 Repeat Exercise 7 with $f(x) = \dfrac{x}{\pi}, -\pi \le x \le \pi$.

Exercise 9 Repeat Exercise 7 with $f(x)$ on $[-\pi, \pi]$ defined by $f(x) = 0$ for $-\pi \le x \le 0$ and $f(x) = \dfrac{x}{\pi}$ for $0 < x \le \pi$.

6.5 Method of Separation of Variables

In this section, we apply the Fourier cosine series result in Theorem 3 of Sect. 6.2, with convergence assured by Theorem 2 of Sect. 6.4 (see Remark 1 on p.304), to solve the "heat diffusion" partial differential equation (PDE) on a perfectly insulated 1-dimensional bounded interval and a 2-dimensional bounded rectangle. It will be seen that our discussion can be easily generalized to s dimensions for $s \geq 2$.

First let us introduce the concept of "separation of variables" and the "principle of superposition". If a function $u(x, t)$ of two variables can be written as

$$u(x, t) = T(t)X(x),$$

where $T(t)$ and $X(x)$ are functions of one variable, then $u(x, t)$ is said to be separable. It is obvious that just about all functions of two variables are not separable. However, by considering all possible separable representations $u(x, t) = T_j(t)X_j(x)$ that satisfy a given (homogeneous) PDE, it is clear that the infinite series

$$u(x, t) = \sum_j b_j T_j(t)X_j(x)$$

also satisfies the same PDE, assuming that the series converges at least in the L_2-norm for certain suitable sequences $\{b_j\}$ of constants. This is called the "principle of superposition". Finally, the sequence $\{b_j\}$ can be determined uniquely using the "initial condition" of the PDE.

In this section, we are only concerned with the heat diffusion PDE

$$\begin{cases} \dfrac{\partial}{\partial t}u(\mathbf{x}, t) = c\nabla^2 u(\mathbf{x}, t), & \mathbf{x} \in D, t \geq 0, \\[2ex] \dfrac{\partial}{\partial \mathbf{n}}u(\mathbf{x}, t) = 0, & \mathbf{x} \in \partial D, \end{cases} \tag{6.5.1}$$

where $c > 0$ is a constant (called thermal conductivity), D is a bounded region in $\mathbb{R}^s$, $s \geq 1$, with piecewise smooth boundary ∂D. Here and throughout this book,

$$\nabla^2 = \frac{\partial^2}{\partial x_1^2} + \cdots + \frac{\partial^2}{\partial x_s^2} \tag{6.5.2}$$

denotes the Laplace operator in the spatial variable $\mathbf{x} = (x_1, \ldots, x_s) \in D$, and $\frac{\partial}{\partial \mathbf{n}}$ denotes the normal derivative with unit normal $\mathbf{n} = \mathbf{n_x}$ for each $\mathbf{x} \in \partial D$.

Recall that the normal derivative is the directional derivative, in the direction of $\mathbf{n}$; namely

$$\frac{\partial}{\partial \mathbf{n}}u(\mathbf{x}, t) = \nabla u(\mathbf{x}, t) \cdot \mathbf{n}, \tag{6.5.3}$$

where ∇u is the gradient of $u(\mathbf{x}, t)$, defined by

$$\nabla u(\mathbf{x}, t) = \left(\frac{\partial}{\partial x_1} u(\mathbf{x}, t), \ldots, \frac{\partial}{\partial x_s} u(\mathbf{x}, t) \right), \tag{6.5.4}$$

and $\nabla u(\mathbf{x}, t) \cdot \mathbf{n}$ is the dot product (or Euclidian inner product) of $\nabla u(\mathbf{x}, t)$ and the unit normal $\mathbf{n}$.

(i) For the 1-dimensional setting, we only consider the closed and bounded interval $D = [0, M]$, $M > 0$, and observe that

$$\nabla^2 u(x, t) = \frac{\partial^2}{\partial x^2} u(x, t),$$

$\partial D = \{0, M\}$, and

$$\frac{\partial}{\partial \mathbf{n}} u(x, t), x \in \partial D$$

becomes $\frac{\partial}{\partial x} u(0, t)$ and $\frac{\partial}{\partial x} u(M, t)$.

(ii) For the 2-dimensional setting, we only focus on the rectangular region $D = [0, M] \times [0, N]$, $M > 0$, $N > 0$, and observe that

$$\nabla^2 u(\mathbf{x}, t) = \nabla^2 u(x, y, t) = \frac{\partial^2}{\partial x^2} u(x, y, t) + \frac{\partial^2}{\partial y^2} u(x, y, t),$$

with $\mathbf{x} = (x, y)$, and

$$\frac{\partial}{\partial \mathbf{n}} u(x, t), \mathbf{x} = (x, y) \in \partial D$$

becomes

$$\frac{\partial}{\partial x} u(x, 0, t), \frac{\partial}{\partial x} u(x, N, t), \frac{\partial}{\partial y} u(0, y, t), \frac{\partial}{\partial y} u(M, y, t). \qquad\blacksquare$$

Returning to the PDE in (6.5.1), we remark that the boundary condition

$$\frac{\partial}{\partial \mathbf{n}} u(\mathbf{x}, t) = 0, \quad \text{for } \mathbf{x} \in \partial D, t \geq 0,$$

in (6.5.1) describes that the "heat content" $u(\mathbf{x}, t)$ for $\mathbf{x} \in D$ is not allowed to diffuse outside D at any time $t \geq 0$. In other words, the region D is perfectly insulated (see Fig. 6.8). Any condition imposed on $\frac{\partial}{\partial \mathbf{n}} u(\mathbf{x}, t)$ is called Neumann's condition. Hence, the PDE described by (6.5.1) is called a Neumann problem, with Neumann's condition equal to zero.

We now turn to the derivation of the general solution of the Neumann (heat) diffusion PDE (6.5.1), by introducing the method of "separation of variables". We will first consider the setting of one spatial variable $\mathbf{x} = x$, then extend it to the
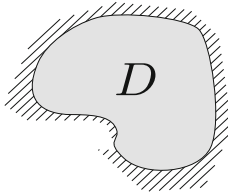
Fig. 6.8 Insulated region

setting of two variables $\mathbf{x} = (x, y)$, and finally conclude with the general s-variable setting.

For $\mathbf{x} = x \in [0, M]$, $M > 0$, the PDE in (6.5.1) becomes

$$
\begin{cases}
\dfrac{\partial}{\partial t} u(x, t) = c \dfrac{\partial^2}{\partial x^2} u(x, t), & x \in [0, M], t \geq 0, \\[4mm]
\dfrac{\partial}{\partial x} u(0, t) = \dfrac{\partial}{\partial x} u(M, t) = 0, & t \geq 0.
\end{cases} \tag{6.5.5}
$$

Hence, if we write

$$
u(x, t) = T(t)X(x),
$$

then (6.5.5) becomes

$$
\begin{cases}
T'(t)X(x) = cT(t)X''(x), & x \in [0, M], t \geq 0, \\
X'(0) = X'(M) = 0.
\end{cases} \tag{6.5.6}
$$

Thus, dividing the PDE in (6.5.6) by $cT(t)X(x)$, we arrive at

$$
\frac{T'(t)}{cT(t)} = \frac{X''(x)}{X(x)},
$$

where the left-hand side is a function independent of the variable x and the right-hand side is a function independent of the variable t. Therefore both sides must be the same constant, say λ, and we now have two ordinary differential equations (ODE), namely:

$$
T'(t) = \lambda c T(t) \tag{6.5.7}
$$

and

$$
\begin{cases}
X''(x) = \lambda X(x), \\
X'(0) = X'(M) = 0.
\end{cases} \tag{6.5.8}
$$

The solution of (6.5.7) is clearly

$$
T(t) = a_0 e^{\lambda c t}
$$

for any constant a_0. To write down the solution of (6.5.8), let us first consider the case $\lambda > 0$, for which the general solution is clearly

$$X(x) = a_1 e^{\sqrt{\lambda}x} + a_2 e^{-\sqrt{\lambda}x}, \quad 0 \le x \le M.$$

For this solution to satisfy the condition $X'(0) = X'(M) = 0$, we have

$$a_1\sqrt{\lambda} - a_2\sqrt{\lambda} = 0, \, a_1\sqrt{\lambda}e^{\sqrt{\lambda}\cdot M} - a_2\sqrt{\lambda}e^{-\sqrt{\lambda}\cdot M} = 0,$$

which implies that $a_1 = a_2 = 0$ (since $\sqrt{\lambda} > 0$). In other words, for (6.5.8) to have a nontrivial solution, the only possibility is $\lambda \le 0$.

For $\lambda = 0$, the ODE in (6.5.8) becomes $X'' = 0$, and hence the general solution is given by $X(x) = a_1 + a_2 x$, where a_1, a_2 are constants. To satisfy the boundary condition $X'(0) = 0$, we have $a_2 = 0$, so that the solution of (6.5.8) is $X(x) = a_1$, for some constant a_1.

Now let us consider $\lambda < 0$, so that $\lambda = -\mu^2$, with $\mu > 0$. Then the general solution of the ODE in (6.5.8) is

$$X(x) = a_1 \cos \mu x + a_2 \sin \mu x,$$

where a_1, a_2 are arbitrary constants. Hence,

$$X'(x) = -a_1 \mu \sin \mu x + a_2 \mu \cos \mu x,$$

so that $X'(0) = a_2 \mu = 0$ implies $a_2 = 0$; that is, $X(x) = a_1 \cos \mu x$ and

$$X'(M) = -a_1 \mu \sin \mu M = 0.$$

Thus, to avoid obtaining the trivial solution $a_1, a_2 = 0$ once again, we must have $\mu = \frac{k\pi}{M}$ for $k = 1, 2, \ldots$. To summarize, the only values of λ in (6.5.7) and (6.5.8) for which nontrivial solutions exist, are

$$\lambda = -\mu^2 = -\left(\frac{k\pi}{M}\right)^2, k = 0, 1, 2, \ldots.$$

We pause for a moment to remark that (6.5.8) is precisely the eigenvalue problem studied in Example 3 of Sect. 2.3 in Chap. 2, but with operator $T = -\frac{d^2}{dx^2}$ instead of $\frac{d^2}{dx^2}$ in (6.5.8). Observe that the eigenvalue $-\left(\frac{k\pi}{M}\right)^2$ is the negative of $\left(\frac{k\pi}{M}\right)^2$ in (2.3.13) on p.95, with $c = M$ in Chap. 2.

Returning to our discussion, since $T(t) = e^{\lambda ct} = e^{-c\left(\frac{k\pi}{M}\right)^2 t}$ and

$$X(x) = \cos \mu x = \cos \frac{k\pi x}{M}$$

(where the reason for not including the coefficients a_0 and a_1 will be clear later), we may conclude that the general solution of the PDE (6.5.5) can be formulated as

$$u(x, t) = \frac{b_0}{2} + \sum_{k=1}^{\infty} b_k e^{-c(\frac{k\pi}{M})^2 t} \cos \frac{k\pi x}{M} \qquad (6.5.9)$$

by the principle of superposition, as described in the beginning of this section. Notice that for $t = 0$, $u(x, 0)$ is a cosine Fourier series in x. Therefore, we have obtained the following result.

Theorem 1 | **Diffusion on insulated bounded interval** | *Let* $u_0 \in L_2[0, M]$, *$M > 0$. Then the solution of the Neumann diffusion PDE (6.5.5) with initial heat content $u(x, 0) = u_0(x)$, $x \in [0, M]$, is given by (6.5.9), with*

$$b_k = \frac{2}{M} \int_0^M u_0(x) \cos \frac{k\pi x}{M} dx, \quad k = 0, 1, 2, \ldots. \qquad (6.5.10)$$

For $u_0 \in L_2[0, M]$, it follows from Theorem 3 of Sect. 6.2 on p.280 (with proof assured by Theorem 2 of Sect. 6.4 on p.302) that the cosine Fourier series

$$u_0(x) = \frac{b_0}{2} + \sum_{k=1}^{\infty} b_k \cos \frac{k\pi x}{M},$$

with the Fourier coefficients $b_0, b_1, \ldots$ given by (6.5.10) (see (6.2.13) of Theorem 3 of Sect. 6.2), converges in the $L_2[0, M]$-norm to $u_0(x)$. Furthermore, if the one-sided derivatives of $u_0(x)$ exist at $x = x_0$, then the series converges to $\left(u_0(x_0^+) + u(x_0^-)\right)/2$, by Theorem 1 of Sect. 6.4 on p.295, provided that $u_0 \in PC[0, M]$. In particular, the series representation (6.5.9) of $u(x, t)$ converges, at least in the $L_2[0, M]$-norm. ∎

Example 1 Solve the heat diffusion PDE

$$\begin{cases} \dfrac{\partial}{\partial t} u(x, t) = c \dfrac{\partial^2}{\partial x^2} u(x, t), & 0 \leq x \leq \pi, t \geq 0, \\[2mm] \dfrac{\partial}{\partial x} u(0, t) = \dfrac{\partial}{\partial x} u(\pi, t) = 0, & t \geq 0, \\[2mm] u(x, 0) = \dfrac{x}{\pi}, & 0 \leq x \leq \pi. \end{cases}$$

Solution We first compute the Fourier cosine series

$$\frac{x}{\pi} = \frac{b_0}{2} + \sum_{k=1}^{\infty} b_k \cos kx,$$

where

$$b_0 = \frac{2}{\pi} \int_0^\pi \frac{x}{\pi} dx = \frac{2}{\pi^2} \cdot \frac{\pi^2}{2} = 1,$$

and, for $k \geq 1$,

$$
\begin{aligned}
b_k &= \frac{2}{\pi} \int_0^\pi \frac{x}{\pi} \cos kx \, dx \\
&= \frac{2}{\pi^2} \left\{ \frac{x \sin kx}{k} \Big|_0^\pi - \int_0^\pi \frac{\sin kx}{k} dx \right\} \\
&= \frac{2}{\pi^2} \{0 - \left(-\frac{\cos kx}{k^2} \right) \Big|_0^\pi \} = \frac{2}{\pi^2 k^2} \left(\cos k\pi - \cos 0 \right) \\
&= \frac{2}{\pi^2 k^2} \left((-1)^k - 1 \right) \\
&= \begin{cases} 0, & \text{for } k = 2j, \\ \frac{2}{\pi^2 (2j-1)^2} \cdot (-2), & \text{for } k = 2j - 1, \end{cases}
\end{aligned}
$$

yielding:

$$\frac{x}{\pi} = \frac{1}{2} - \frac{4}{\pi^2} \sum_{j=1}^{\infty} \frac{1}{(2j-1)^2} \cos \left((2j-1)x \right), \quad 0 \leq x \leq \pi.$$

Therefore, the solution of the required heat diffusion PDE is given by

$$u(x, t) = \frac{1}{2} - \frac{4}{\pi^2} \sum_{j=1}^{\infty} \frac{1}{(2j-1)^2} e^{-c(2j-1)^2 t} \cos \left((2j-1)x \right),$$

where $0 \leq x \leq \pi, t \geq 0$.   ∎

Example 2 Solve the heat diffusion PDE

$$
\begin{cases}
\dfrac{\partial}{\partial t} u(x, t) = c \dfrac{\partial^2}{\partial x^2} u(x, t), & 0 \leq x \leq \pi, t \geq 0, \\[2mm]
\dfrac{\partial}{\partial x} u(0, t) = \dfrac{\partial}{\partial x} u(\pi, t) = 0, & t \geq 0, \\[2mm]
u(x, 0) = 1 + \cos 2x, & 0 \leq x \leq \pi.
\end{cases}
$$

Solution Since the initial function

$$u_0(x) = 1 + \cos 2x$$

is already a Fourier cosine series, with $b_0 = 2, b_2 = 1, b_k = 0$ for $k \neq 0, 2$, the solution of the required heat diffusion PDE is simply:

$$u(x, t) = \frac{2}{2} + 0 \cdot e^{-c(\frac{1 \cdot \pi}{\pi})^2 t} \cos\left(\frac{1 \cdot \pi}{\pi}x\right) + 1 \cdot e^{-c(\frac{2 \cdot \pi}{\pi})^2 t} \cos\left(\frac{2 \cdot \pi}{\pi}x\right)$$

$$+ 0 \cdot e^{-c(\frac{3 \cdot \pi}{\pi})^2 t} \cos\left(\frac{3 \cdot \pi}{\pi}x\right) + \cdots$$

$$= 1 + e^{-c4t} \cos 2x.$$

To verify this answer, we observe that

$$\left.\begin{array}{l}
\text{(i) } \dfrac{\partial}{\partial t}u(x, t) = -4ce^{-4ct} \cos 2x \\[2mm]
\text{(ii) } \dfrac{\partial^2}{\partial x^2}u(x, t) = 4e^{-4ct}(-\cos 2x)
\end{array}\right\} \implies \frac{\partial}{\partial t}u(x, t) = c\frac{\partial^2}{\partial x^2}u(x, t),$$

$$\text{(iii) } \frac{\partial}{\partial x}u(x, t) = -2e^{-4ct} \sin 2x \implies$$

$$\frac{\partial}{\partial x}u(0, t) = -2e^{-4ct} \sin 0 = 0, \quad \frac{\partial}{\partial x}u(\pi, t) = -2e^{-4ct} \sin 2\pi = 0.$$

Finally, $u(x, 0) = 1 + e^0 \cos 2x = 1 + \cos 2x = u_0(x)$. ∎

The method of separation of variables extends to any bounded region $D \subset \mathbb{R}^s$, $s > 1$, as follows. We begin by writing

$$u(\mathbf{x}, t) = T(t)U(\mathbf{x}), \mathbf{x} \in D.$$

Then, for $u(\mathbf{x}, t)$ in the PDE in (6.5.1), we have

$$T'(t)U(x) = cT(t)\nabla^2 U(\mathbf{x})$$

so that

$$\frac{T'(t)}{cT(t)} = \frac{\nabla^2 U(\mathbf{x})}{U(\mathbf{x})},$$

which must be a constant λ (since the left-hand side is independent of the variable $\mathbf{x} \in D$ and the right-hand side is a function independent of the variable $t > 0$). Hence, we have one ODE given by

$$T'(t) - \lambda cT(t) = 0, \tag{6.5.11}$$

which is the same as (6.5.7) and one ODE given by

$$\nabla^2 U(\mathbf{x}) = \lambda U(\mathbf{x}), \tag{6.5.12}$$

which is an eigenvalue problem. Since

$$T(t) = a_0 e^{\lambda c t},$$

as in the 1-dimensional diffusion problem, and by keeping in mind that the PDE in (6.5.1) describes a diffusion process, we may conclude that $\lambda \leq 0$, as discovered earlier. Unfortunately, for an arbitrary region $D \subset \mathbb{R}^s$, $s \geq 2$, the eigenvalue problem

$$\nabla^2 U(\mathbf{x}) = -\mu^2 U(\mathbf{x}), \mathbf{x} \in D$$

(with $\lambda = -\mu^2$, $\mu \geq 0$), does not have an explicit solution (see the discussion on numerical approximation via the Rayleigh quotient with $T = -\nabla^2$ in Sect. 2.3 of Chap. 2). To apply the Fourier series method, we only consider

$$D = [0, M_1] \times \cdots \times [0, M_s], \tag{6.5.13}$$

where $M_1, \ldots, M_s > 0$. Let us first discuss the case $s = 2$, and let $M = M_1$, $N = M_2$ for convenience. Then, by setting $x = x_1$ and $y = x_2$, the Neumann diffusion PDE (6.5.1) becomes

$$\begin{cases} \dfrac{\partial}{\partial t} u(x, y, t) = c\nabla^2 u(x, y, t), & x \in [0, M], y \in [0, N], t \geq 0, \\[2mm] \dfrac{\partial}{\partial x} u(0, y, t) = 0, \dfrac{\partial}{\partial x} u(M, y, t) = 0, & y \in [0, N], t \geq 0, \\[2mm] \dfrac{\partial}{\partial y} u(x, 0, t) = 0, \dfrac{\partial}{\partial y} u(x, N, t) = 0, & x \in [0, M], t \geq 0. \end{cases} \tag{6.5.14}$$

Then for

$$U(x, y) = X(x)Y(y),$$

we have

$$\nabla^2 U(x, y) = X''(x)Y(y) + X(x)Y''(y),$$

and hence $\nabla^2 U = \lambda U$ becomes

$$X''(x)Y(y) + X(x)Y''(y) = \lambda X(x)Y(y),$$

or equivalently,

$$\frac{X''(x)}{X(x)} = -\frac{Y''(y)}{Y(y)} + \lambda = \widetilde{\lambda}$$

for some constant $\widetilde{\lambda}$ (since $-Y''(y)/Y(y) + \lambda$ is independent of x and $X''(x)/X(x)$ is independent of y). Now, by using the two-point boundary condition

$$X'(0) = X'(M) = 0,$$

it can be shown that $\widetilde{\lambda} = -(k\pi/M)^2$ for $k = 0, 1, 2, \ldots$, as in the setting of one spatial variable, so that

$$X(x) = \text{constant mulitple of } \cos\frac{k\pi x}{M}.$$

Similarly, for $Y''(y)/Y(y) = \lambda - \widetilde{\lambda}$, the two-point boundary condition

$$Y'(0) = Y'(N) = 0,$$

implies that $\lambda - \widetilde{\lambda} = -(\ell\pi/N)^2$ for $\ell = 0, 1, 2, \ldots$, so that

$$\lambda = \widetilde{\lambda} - \left(\frac{\ell\pi}{N}\right)^2 = -\left(\frac{k\pi}{M}\right)^2 - \left(\frac{\ell\pi}{N}\right)^2,$$

which is ≤ 0, as expected in view of the diffusion process described by the PDE. Also,

$$Y(y) = \text{constant mulitple of } \cos\frac{\ell\pi x}{N}.$$

Hence, by the principle of superposition, we have

$$u(x, y, t) = \sum_{k=0}^{\infty}\sum_{\ell=0}^{\infty} d(k, \ell)b_{k,\ell}e^{-c\left(\left(\frac{k\pi}{M}\right)^2 + \left(\frac{\ell\pi}{N}\right)^2\right)t} \cos\left(\frac{k\pi}{M}x\right)\cos\left(\frac{\ell\pi}{N}y\right),$$

$$(6.5.15)$$

where we have introduced the notation

$$d(k, \ell) = 2^{-(\delta_k + \delta_\ell)}, k, \ell = 0, 1, 2, \ldots, \qquad (6.5.16)$$

by using the Kronecker delta symbol δ_j (for $j = k$ and $j = \ell$), and where $b_{k,\ell}$ are certain constants. $u(x, y, t)$ in (6.5.15) gives the general solution of the PDE (6.5.14) before the initial condition is considered. To impose the initial condition, we have

$$u_0(x, y) = u(x, y, 0) = \sum_{k=0}^{\infty}\sum_{\ell=0}^{\infty} d(k, \ell)b_{k,\ell}\cos\left(\frac{k\pi}{M}x\right)\cos\left(\frac{\ell\pi}{N}y\right), \quad (6.5.17)$$

which is the Fourier cosine series representation of $u_0(x, y)$. Thus, $b_{k,\ell}, k = 0, 1, \ldots, \ell = 0, 1, \ldots$ in (6.5.15) are the coefficients of the cosine series of the initial function $u_0(x, y)$, namely:

$$b_{k\ell} = \frac{4}{MN} \int_0^M \int_0^N u_0(x, y) \cos\left(\frac{k\pi}{M}x\right) \cos\left(\frac{\ell\pi}{N}y\right) dxdy, \qquad (6.5.18)$$

for $k, \ell = 0, 1, 2, \ldots$.   ∎

For $s \geq 3$, we may consider

$$U(\mathbf{x}) = U(x_1, \ldots x_{s-1}, x_s) = V(x_1, \ldots, x_{s-1})X(x_s),$$

so that

$$\nabla_s^2 U = X\nabla_{s-1}^2 V + X''V$$

(where the subscript of ∇^2 denotes the dimension of the Laplace operator). Hence, by an induction argument, we have the following result, which extends Theorem 1 to any dimension $s \geq 2$.

For convenience, the notation $d(k, \ell)$ in (6.5.16) is extended to an arbitrary $s \geq 2$; namely

$$d(k_1, \ldots, k_s) = 2^{-\sum_{j=1}^s \delta_{k_j}}.$$

Theorem 2 | **Diffusion in higher dimensions** | *Let $u_0 \in L_1([0, M_1] \times \cdots \times [0, M_s])$, where $M_1, \ldots, M_s > 0$. Then the solution of the Neumann diffusion PDE (6.5.1) for $D = [0, M_1] \times \cdots \times [0, M_s]$ with initial heat content $u(\mathbf{x}, 0) = u_0(\mathbf{x}), \mathbf{x} \in D$, is given by*

$$u(\mathbf{x}, t) =$$

$$\sum_{k_1=0}^\infty \cdots \sum_{k_s=0}^\infty d(k_1, \ldots, k_s) b_{k_1\ldots k_s} e^{-c\sum_{j=1}^s \left(\frac{k_j\pi}{M_j}\right)^2 t} \cos\left(\frac{k_1\pi}{M_1}x_1\right) \ldots \cos\left(\frac{k_s\pi}{M_s}x_s\right),$$
$$(6.5.19)$$

with

$$b_{k_1\ldots k_s} = \frac{2^s}{M_1 \ldots M_s} \int_D u_0(\mathbf{x}) \cos\left(\frac{k_1\pi}{M_1}x_1\right) \ldots \cos\left(\frac{k_s\pi}{M_s}x_s\right) d\mathbf{x},$$

where the convergence of the series in (6.5.19) is in the $L_2(D)$-norm.

Example 3 Solve the Neumann diffusion PDE (6.5.14) for $M = N = \pi$ with initial heat content

$$u(\mathbf{x}, 0) = u_0(\mathbf{x}) = 1 + 2\cos x \cos y + \cos x \cos 3y.$$

Solution Since the representation of $u_0(x, y)$ is already its Fourier cosine series representation, it follows from Theorem 2 that the solution is given by

$$u(x, y, t) = 1 + 2e^{-c(1^2+1^2)t} \cos x \cos y + e^{-c(1^2+3^2)t} \cos x \cos 3y$$
$$= 1 + 2e^{-2ct} \cos x \cos y + e^{-10ct} \cos x \cos 3y.$$

■

Exercises

Exercise 1 Solve the heat diffusion PDE (6.5.5) for $c = 2$, $M = \pi$, and

$$u(x, 0) = u_0(x) = |x - \frac{\pi}{2}|, \quad 0 \le x \le \pi.$$

Exercise 2 Apply the method of "separation of variables" to separate the following PDE's into ODE's.

(a) $\dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2} = 0.$

(b) $\dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial u}{\partial x} + \dfrac{\partial u}{\partial y} = 0.$

(c) $\dfrac{\partial^2 u}{\partial x^2} + f(x)\dfrac{\partial u}{\partial x} + \dfrac{\partial^2 u}{\partial y^2} = 0.$

Exercise 3 Solve the heat diffusion PDE (6.5.14) for $c = 3$, $M = N = \pi$, and

$$u_0(x, y) = 2 + \cos 2x \cos 3y + 4 \cos 5x \cos 2y.$$

Exercise 4 Repeat Exercise 3 for

$$u_0(x, y) = \sum_{k=0}^{10} \sum_{j=0}^{5} \frac{k}{j+1} \cos kx \cos jy.$$

Exercise 5 Repeat Exercise 3 for

$$u_0(x, y) = |x - \frac{\pi}{2}| + |y - \frac{\pi}{2}|, \quad 0 \le x, y \le \pi.$$

Exercise 6 Repeat Exercise 3 for

$$u_0(x, y) = |(x - \frac{\pi}{2})(y - \frac{\pi}{2})|, \quad 0 \le x, y \le \pi.$$

Exercise 7 Apply the method of "separation of variables" to solve the 1-dimensional PDE of "vibrating string" with no movement at the two end-points, given by

$$
\begin{cases}
\dfrac{\partial^2}{\partial t^2} u(x, t) = c \dfrac{\partial^2}{\partial x^2} u(x, t), & x \in [0, M], t \geq 0, \\[4mm]
u(0, t) = u(M, t) = 0, & t \geq 0,
\end{cases}
$$

where $c > 0$ is a constant.

Exercise 8 Apply the method of "separation of variables" to separate the PDE of the "wave equation of the two spatial variables" for $u = u(x, y, t)$, given by

$$
\frac{\partial^2 u}{\partial t^2} = c^2 \nabla^2 u, \quad c > 0 \text{ constant},
$$

where $\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$, into a PDE in the space variables x, y and an ODE in the time variable t.

Exercise 9 The PDE model of a "vibrating membrane" is the wave equation with zero boundary condition. As a continuation of Exercise 8, solve the following PDE that describes a vibrating membrane on $[0, M] \times [0, N]$ with the given Fourier sine series representation of the initial vibration:

$$
\begin{cases}
\frac{\partial^2}{\partial t^2} u(x, y, t) = c^2 \nabla^2 u(x, y, t), & x \in [0, M], y \in [0, N], t \geq 0, \\[2mm]
u(x, y, 0) = 0, & x \in [0, M], y \in [0, N], \\[2mm]
\frac{\partial}{\partial t} u(x, y, 0) = \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} b_{m,n} \sin \frac{m \pi x}{M} \sin \frac{n \pi y}{N}, & x \in [0, M], y \in [0, N],
\end{cases}
$$

where $\displaystyle\sum_{m=0}^{\infty} \sum_{n=0}^{\infty} |b_{m,n}| < \infty$.

Chapter 7
Fourier Time-Frequency Methods
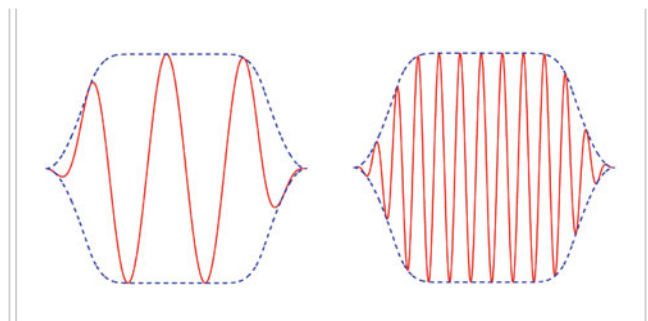
When non-periodic functions $f \in PC(-\infty, \infty)$ are considered as analog signals with time-domain $(-\infty, \infty)$, the notion of the Fourier transform (FT)

$$\widehat{f}(\omega) = \int_{-\infty}^{\infty} f(x)e^{-ix\omega}\, dx$$

is introduced in Sect. 7.1 to reveal the frequency content of $f \in L_2(\mathbb{R})$, where the most useful properties of the FT are derived. In addition, the FT of the Gaussian function is computed, and it will be clear that the FT of the Gaussian function is again another Gaussian. Analogous to Fejér's kernels introduced in Sect. 6.3 of the previous chapter, it is shown that the family of Gaussian functions,

$$g_\sigma(x) = \frac{1}{2\sigma\sqrt{\pi}} e^{-\frac{1}{4\sigma^2}x^2},$$

with parameter σ, also constitutes a positive approximate identity, but now for the entire real line $\mathbb{R}$, as compared with the interval $[-\pi, \pi]$ for the sequence of Fejér's kernels.

The Fourier transform defined for functions $f \in L_1(\mathbb{R})$ in Sect. 7.1 is extended to $L_2(\mathbb{R})$ in Sect. 7.2, where Plancherel's formula (also called Parseval's identity to match that of the Fourier series established in Sect. 6.1 of Chap. 6) is derived by applying the property of positive approximate identity of the Gaussian function g_σ and the Gaussian formulation of the FT of g_σ. Under the assumption that the Fourier transform $\widehat{f}$ of the given function f is also in $L_1(\mathbb{R})$, it is shown in this section that f can be recovered from $\widehat{f}(\omega)$ by the inverse Fourier transform (IFT), defined by

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \widehat{f}(\omega)e^{i\omega x}\, d\omega.$$

For the time-domain $(-\infty, \infty)$, convolution of two functions $f(x)$ and $h(x)$ is defined by

$$(f * h)(x) = \int_{-\infty}^{\infty} f(t)h(x - t)dt,$$

and one of the most important properties of the FT is that it maps convolutions to products, namely:

$$\widehat{(f * h)}(\omega) = \widehat{f}(\omega)\widehat{h}(\omega).$$

Hence, for applications to signal processing, by selecting an appropriate filter function $h(x)$, the convolution operation on an analog signal $f(x)$ results in modulation of its frequency content $\widehat{f}(\omega)$ with the desired filter characteristic $\widehat{h}(\omega)$ in the frequency-domain. As an application to ideal lowpass filtering, we will derive the sampling theorem at the end of Sect. 7.2.

Section 7.3 can be considered as a continuation of Sect. 6.5 on the solution of the diffusion PDE, but now for the entire Euclidean space $\mathbb{R}^s$ instead of bounded rectangular regions in Sect. 6.5. Since it is not a boundary value problem, we may now consider the diffusion PDE as an initial value problem, with initial heat content given by a function u_0 defined on $\mathbb{R}^s$ as the initial condition. For $\mathbb{R}^1$, it will be shown in this section that when the square of the parameter σ of the family of Gaussian functions g_σ is replaced by ct, where $c > 0$ is a constant and t denotes the time parameter, then spatial convolution of the initial heat content u_0 with g_σ yields the solution of the initial value heat diffusion PDE, with the positive constant c being the thermal conductivity of the diffusion media. This derivation is extended to $\mathbb{R}^s$, for any spatial dimension $s \geq 2$. Observe that since the Gaussian function is well localized, the initial function u_0 is allowed to be of polynomial growth as well as being a periodic function, such as a Fourier cosine or sine series. In fact, for the region $\mathbb{R}^2$, when a Fourier cosine series defined on the rectangular region $[0, M] \times [0, N]$ is considered as the initial condition for the diffusion PDE for the entire spatial domain $\mathbb{R}^2$, it is shown in the final example, Example 6, of this section that the solution is the same as that of the diffusion PDE for the perfectly insulated bounded region $[0, M] \times [0, N]$, with zero Neumann condition, obtained by applying the method of separation of variables.

In many applications, it is desirable to localize both in time and in space. The choice of filter functions $h(x)$ to simultaneously localize an analog signal in both the time-domain and the frequency-domain is studied in Sect. 7.4, where an identity for simultaneous time-frequency localization is established. In addition, the general formulation of the Gaussian

$$g_{c,\gamma,b}(x) = ce^{-\gamma(x-b)^2}$$

with $\gamma > 0, c \neq 0$ and $b \in \mathbb{R}$, is shown to provide the optimal time-frequency localization window function, as governed by the uncertainty principle, also to be derived in Sect. 7.4. If a time-window function $u(x)$ is used to define the localized

(or short-time) Fourier transform $(\mathbb{F}_u f)(x, \omega)$ of any $f \in L_2(\mathbb{R})$, then under the conditions that both $u(x)$ and its Fourier transform $\widehat{u}(\omega)$ are in $(L_1 \cap L_2)(\mathbb{R})$ and that $u(0) \neq 0$, it will be shown also in Sect. 7.4 that $f(x)$ can be recovered from $(\mathbb{F}_u f)(x, \omega)$ by applying the inverse FT operation.

On the other hand, for practical applications, the time-frequency coordinates (x, ω) are sampled at $(x, \omega) = (ma, 2\pi kb)$, where $a, b > 0$ are fixed and m, k run over the set $\mathbb{Z}$ of all integers. Then the localized Fourier transform $(\mathbb{F}_u f)(x, \omega)$, introduced in Sect. 7.4, at these sample points can be expressed as an inner product, namely: $(\mathbb{F}_u f)(ma, 2\pi kb) = \langle f, h_{ma,kb} \rangle$, where $h_{ma,kb}(x) = u(x - ma)e^{i2\pi kbx}$, $m, k \in \mathbb{Z}$, could be viewed as some time-frequency "basis" functions, with m running over the discrete time-domain $\mathbb{Z}$, and n running over the frequency domain $\mathbb{Z}$. Unfortunately, for good time-frequency localization (that is, for u to satisfy $\Delta_u \Delta_{\widehat{u}} < \infty$), it is necessary to sample at $(ma, 2\pi kb)$ with $ab < 1$ for the existence of such window functions $u(x)$. This is the so-called Balian-Low constraint. On the other hand, the choice of $a = b = 1$ significantly facilitates computational efficiency. The good news is that the Balian-Low constraint can be removed by replacing $e^{i2\pi kx}$ (for $a = b = 1$) with cosine and/or sine functions. The study of such time-frequency bases will be studied in Sect. 7.5.

For the book to be self-contained, a discussion of the integration theory from Real Analysis, including Fubini's theorem, Minkowski's integral inequality, Fatou's lemma, Lebesgue's dominated convergence theorem, and Lebesgue's integration theorem, is presented in the last section to facilitate the derivation of certain theorems in this chapter.

7.1 Fourier Transform

The discrete Fourier transform (DFT) introduced in Chap. 4 allows us to study the frequency behavior of digital data. More precisely, when $\mathbf{x} = (x_0, \ldots, x_{n-1}) \in \mathbb{C}^n$ is considered as a digital signal with time-domain $(0, \ldots, n-1)$, then its n-point DFT

$$\widehat{\mathbf{x}} = (\widehat{x}_0, \ldots, \widehat{x}_{n-1}) = \mathbb{F}_n \mathbf{x}$$

reveals the frequency content of the digital signal $\mathbf{x}$ (see Example 2 on p.176).

In Sect. 6.1, the notion of Fourier coefficients is introduced to study the frequency behavior of analog signals with time-domain given by a bounded interval $[a, b]$, where $-\infty < a < b < \infty$. By a simple change of variables as discussed in Sect. 6.1, we may replace $[a, b]$ by $[-\pi, \pi]$. Hence, by periodic extension of functions in $PC[-\pi, \pi]$ to $PC^\star_{2\pi}$, the Fourier series can be considered as the "inverse" transform of the sequence of Fourier coefficients to recover the analog signal. In this section, we introduce the notion of the Fourier transform to study functions in $PC(-\infty, \infty)$. Hence, the Fourier transform of an analog signal with time-domain $(-\infty, \infty)$ reveals its entire frequency content, and the inverse Fourier transform to be studied in Sect. 7.2 recovers the analog signal from the frequency content.

Definition 1 | **Fourier transform** | *Let $f \in L_1(\mathbb{R})$. The Fourier transform (FT) of f, denoted by $\widehat{f}$ or $\mathbb{F}f$, is defined by*

$$\widehat{f}(\omega) = (\mathbb{F}f)(\omega) = \int_{-\infty}^{\infty} f(x)e^{-i\omega x}\, dx, \quad \omega \in \mathbb{R}. \tag{7.1.1}$$

Sometimes it may be more convenient to write $\widehat{f}(\omega)$ as $(f)^{\wedge}(\omega)$. See, for example, (7.1.10)–(7.1.12).

Observe that since $e^{-ix\omega} = \cos x\omega - i \sin x\omega$, the FT of $f(x)$ reveals the frequency content

$$\widehat{f}(\omega) = \int_{-\infty}^{\infty} f(x) \cos \omega x\, dx - i \int_{-\infty}^{\infty} f(x) \sin \omega x\, dx$$

of $f(x)$ in terms of the oscillation of the cosine and sine functions, with frequency defined by $\omega/2\pi$ Hz.

Example 1 *Compute the Fourier transform $\widehat{f}(\omega)$ of $f(x) = e^{-a|x|}$, where $a > 0$.*

Solution Clearly $f \in L_1(\mathbb{R})$. Since $f(x)$ is even,

$$\begin{aligned}
\widehat{f}(\omega) &= \int_{-\infty}^{\infty} e^{-a|x|}e^{-i\omega x}dx = 2 \int_{0}^{\infty} e^{-ax} \cos \omega x dx \\
&= 2\left[\frac{e^{-ax}}{\omega^2 + a^2}(\omega \sin \omega x - a \cos \omega x)\right]_0^{\infty} = \frac{2a}{\omega^2 + a^2}.
\end{aligned}$$

∎

The Fourier transform $\widehat{f} = \mathbb{F}f$ of a function $f \in L_1(\mathbb{R})$ enjoys the following properties:

Theorem 1 *Let $f \in L_1(\mathbb{R})$ with Fourier transform $\widehat{f}(\omega)$. Then*

 (i) *$\widehat{f} \in L_\infty(\mathbb{R})$ with $\|\widehat{f}\|_\infty \le \|f\|_1$.*
 (ii) *$\widehat{f}(\omega)$ is uniformly continuous on $\mathbb{R}$.*
 (iii) *$\widehat{f}(\omega) \to 0$ as $|\omega| \to \infty$.*
 (iv) *If $f'(x)$ exists on $\mathbb{R}$ and $f' \in L_1(\mathbb{R})$, then $(\mathbb{F}f')(\omega) = i\omega \widehat{f}(\omega)$.*
 (v) *If $x^k f(x) \in L_1(\mathbb{R})$ for some $k \ge 1$, then $\widehat{f} \in C^k(\mathbb{R})$ with*

$$\frac{d^k}{d\omega^k} \widehat{f}(\omega) = (-i)^k \int_{-\infty}^{\infty} x^k f(x)e^{-i\omega x}dx, \ \omega \in \mathbb{R}. \tag{7.1.2}$$

Since the proof of (i) is straightforward, it is safe to leave it as an exercise (see Exercise 2). To prove (ii), we observe that for any $\epsilon > 0$,

$$\sup_{\omega} \left| \widehat{f}(\omega + \epsilon) - \widehat{f}(\omega) \right| = \sup_{\omega} \left| \int_{-\infty}^{\infty} f(x) \left(e^{-i\epsilon x} - 1 \right) e^{-i\omega x} \, dx \right|$$

$$\leq \int_{-\infty}^{\infty} |f(x)| \left| e^{-i\epsilon x} - 1 \right| dx \to 0,$$

for $\epsilon \to 0$, since

$$|f(x)| \left| e^{-i\epsilon x} - 1 \right| \leq 2|f(x)|$$

for any ϵ and $f(x) \in L_1(\mathbb{R})$, so that the "Lebesgue's dominated convergence theorem" from Real Analysis (see Theorem 4 in the Appendix of this chapter) can be applied to interchange limit and integration.

Property (iii) is called the "Riemann-Lebesgue lemma". To prove (iii), we may assume that f is a compactly supported piecewise constant function, namely: $f(x) = \sum_{k=1}^{n} c_k \chi_{[a_{k-1}, a_k)}(x)$. The reason is that an $L_1(\mathbb{R})$ function can be approximated arbitrarily well by such functions, with Fourier transform that satisfies (i). Then

$$\widehat{f}(\omega) = \sum_{k=1}^{n} c_k \int_{a_{k-1}}^{a_k} e^{-i\omega x} dx = \frac{1}{-i\omega} \sum_{k=1}^{n} c_k (e^{-i\omega a_k} - e^{-i\omega a_{k-1}}) \to 0,$$

as $\omega \to \infty$.

The proof of (iv) follows from "integration by parts" which can be shown rigorously, under the assumption that $f' \in L_1(\mathbb{R})$ and by observing that $f(x) \to 0$ as $|x| \to \infty$.

For (v), we first observe that if $f(x) \in L_1(\mathbb{R})$ and $x^k f(x) \in L_1(\mathbb{R})$ as well, then $x^j f(x) \in L_1(\mathbb{R})$ for $1 \leq j \leq k - 1$. The property $\widehat{f} \in C^k(\mathbb{R})$ can be verified iteratively by showing that $\widehat{f} \in C^j(\mathbb{R})$, with $j = 1, \ldots, k$. Here, we may consider only the case $k = 1$, namely:

$$\frac{d}{d\omega} \widehat{f}(\omega) = \lim_{\xi \to 0} \frac{\widehat{f}(\omega + \xi) - \widehat{f}(\omega)}{\xi}$$

$$= \lim_{\xi \to 0} \int_{-\infty}^{\infty} f(x) \frac{e^{-i(\omega + \xi)x} - e^{-i\omega x}}{\xi} dx$$

$$= \int_{-\infty}^{\infty} f(x) e^{-i\omega x} \lim_{\xi \to 0} \frac{e^{-i\xi x} - 1}{\xi} dx$$

$$= \int_{-\infty}^{\infty} f(x)(-ix) e^{-i\omega x} dx,$$

where the interchange of limit and integration in the third equality is allowed by applying Lebesgue's dominated convergence theorem to the function

$$q(x, \xi) = f(x) \frac{e^{-i(\omega + \xi)x} - e^{-i\omega x}}{\xi} = f(x) e^{-i\omega x} \frac{e^{-i\xi x} - 1}{\xi},$$

since $|q(x, \xi)| \leq |x| |f(x)|$ for any real ξ (see Exercise 14); that is, $q(x, \xi)$ is dominated by an $L_1(\mathbb{R})$ function $|x| |f(x)|$.

The continuity of $\frac{d}{d\omega} \widehat{f}(\omega)$ follows from the expression of $\frac{d}{d\omega} \widehat{f}(\omega)$ in (7.1.2) and (ii) with $f(x)$ replaced by $xf(x)$.   ∎

In the following, we list some important operations of functions in $L_1(\mathbb{R})$ or $L_1[0, \infty)$.

(a) **Even extension.** Let $f \in L_1[0, \infty)$. The even extension $f_e(x)$ of $f(x)$ is defined by $f_e(x) = f(x)$ for $x \geq 0$ and

$$f_e(x) = f(-x), \quad \text{for } x < 0. \tag{7.1.3}$$

Note that $f_e \in L_1(\mathbb{R})$.

(b) **Translation.** Let $f \in L_1(\mathbb{R})$ and $b \in \mathbb{R}$. The translation operator T_b is defined by

$$(T_b f)(x) = f(x - b). \tag{7.1.4}$$

(c) **Dilation.** Let $f \in L_1(\mathbb{R})$ and $a > 0$. The dilation operator D_a is defined by

$$(D_a f)(x) = f(ax). \tag{7.1.5}$$

Note that $D_a f \in L_1(\mathbb{R})$.

(d) **Frequency modulation.** Let $f \in L_1(\mathbb{R})$ and $0 \neq c \in \mathbb{R}$. Then the (frequency) modulation operator M_c is defined by

$$(M_c f)(x) = f(x)e^{icx}. \tag{7.1.6}$$

Note that $|(M_c f)(x)| = |f(x)|$.

(e) **Convolution.** Let f be a function on $\mathbb{R}$. Then for a given "filter", which is another function on $\mathbb{R}$, the convolution of $f(x)$ with the filter function $h(x)$ is defined by

$$(f * h)(x) = \int_{-\infty}^{\infty} f(t)h(x - t)dt. \tag{7.1.7}$$

Remark 1 By applying the Cauchy-Schwarz inequality in Theorem 3 on p.27 (or Hölder's inequality for $p = q = 2$ in (1.2.21) on p.22), it is clear that $f * h \in L_\infty(\mathbb{R})$ provided that $f, h \in L_2(\mathbb{R})$ (see Exercise 4). If $f, h \in L_1(\mathbb{R})$, then $f * h \in L_1(\mathbb{R})$ by applying Fubini's theorem in Theorem 1 of the Appendix (see Exercise 5). More general, by the Minkowski inequality for integrals, it can be shown that for $1 \leq p, q < \infty$,

$$\|f * h\|_r \leq \|f\|_p \|h\|_q,$$

where $r > 0$ is given by $\frac{1}{r} + 1 = \frac{1}{p} + \frac{1}{q}$.   ∎

Remark 2 By the change of variables of integration $u = x - t$, we see that

$$f * h = h * f.$$

Thus, the convolution operator defined by (7.1.7) is commutative.

Observe that the convolution operator $*$ defined in (6.3.3) of Definition 1 on p. 288 for $PC_{2\pi}^{\star}$ differs from that in (7.1.7) for $L_2(\mathbb{R})$ only in the limits of integration.

∎

Next we derive the formulas for the Fourier transform of functions resulting from the operations introduced in (7.1.3)–(7.1.7).

Theorem 2 | **Properties of Fourier transform**

(i) *Let f_e be the even extension of $f \in L_1[0, \infty)$. Then*

$$\widehat{f_e}(\omega) = 2 \int_0^\infty f(x) \cos \omega x \, dx. \tag{7.1.8}$$

(ii) *Let $f \in L_1(\mathbb{R})$ and $b \in \mathbb{R}$. Then*

$$(T_b f)^{\wedge}(\omega) = e^{-ib\omega} \widehat{f}(\omega); \tag{7.1.9}$$

that is, $\mathbb{F}T_b = M_{-b}\mathbb{F}$.

(iii) *Let $f \in L_1(\mathbb{R})$ and $a \neq 0$. Then*

$$(D_a f)^{\wedge}(\omega) = \frac{1}{a} \widehat{f}\left(\frac{\omega}{a}\right); \tag{7.1.10}$$

that is, $\mathbb{F}D_a = \frac{1}{a} D_{a^{-1}}\mathbb{F}$.

(iv) *Let $f \in L_1(\mathbb{R})$ and $0 \neq c \in \mathbb{R}$. Then*

$$(M_c f)^{\wedge}(\omega) = \widehat{f}(\omega - c); \tag{7.1.11}$$

that is, $\mathbb{F}M_c = T_c\mathbb{F}$.

(v) *Let $f, h \in L_1(\mathbb{R})$. Then $f * h \in L_1(\mathbb{R})$ and*

$$(f * h)^{\wedge}(\omega) = \widehat{f}(\omega)\widehat{h}(\omega). \tag{7.1.12}$$

Proofs of (7.1.8)–(7.1.11) are straightforward (see Exercises 15–18), and derivation of (7.1.12) is a consequence of interchanging the order of integrations by applying Fubini's theorem, as follows.

$$
\begin{aligned}
(f * h)^\wedge(\omega) &= \int_{-\infty}^{\infty} e^{-i\omega x} \left\{ \int_{-\infty}^{\infty} f(t) h(x - t) dt \right\} dx \\
&= \int_{-\infty}^{\infty} f(t) \left\{ \int_{-\infty}^{\infty} e^{-i\omega x} h(x - t) dx \right\} dt \\
&= \int_{-\infty}^{\infty} f(t) \left\{ \int_{-\infty}^{\infty} e^{-i\omega(y+t)} h(y) dy \right\} dt \\
&= \int_{-\infty}^{\infty} f(t) e^{-i\omega t} \widehat{h}(\omega) dt \\
&= \widehat{f}(\omega) \widehat{h}(\omega).
\end{aligned}
$$

∎

In the rest of this section, we consider the Gaussian function

$$
g_\sigma(x) = \frac{1}{2\sigma\sqrt{\pi}} e^{-\left(\frac{x}{2\sigma}\right)^2}, \tag{7.1.13}
$$

where $\sigma > 0$.

Remark 3   When the Gaussian function g_σ is used as the time localization window function u in (7.4.1) on p.351, the radius of this window function g_σ is shown to be $\Delta_{g_\sigma} = \sigma$ in Theorem 2 of Sect. 7.4 on p.355. In addition, since the Fourier transform of g_σ is simply $\widehat{g}_\sigma(\omega) = e^{-\sigma^2\omega^2}$, as shown in Theorem 3, the radius of the frequency localization window function $\widehat{g}_\sigma$ is $\dfrac{1}{2\sigma}$ as shown in Theorem 2 of Sect. 7.4. On the other hand, in the Statistics literature, the Gaussian function

$$
f_\sigma(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{x}{\sqrt{2}\sigma}\right)^2}
$$

is often adopted, since σ is the "standard derivation" of the probability distribution function $y = f_\sigma(x)$, with "variance" σ^2. Hence, in view of

$$
f_\sigma(x) = g_{\sigma/\sqrt{2}}(x),
$$

it is clear that when g_σ is used as a probability distribution function, its variance is

$$
\mathrm{var}(g_\sigma) = 2\sigma^2.
$$

∎

The division by $2\sigma\sqrt{\pi}$ in (7.1.13) is to assure the integral to be equal to 1, namely:

$$
\int_{-\infty}^{\infty} g_\sigma(x) dx = 1, \ \text{all} \ \sigma > 0. \tag{7.1.14}
$$

To prove (7.1.14), observe that by changing the Cartesian coordinates to polar coordinates, we have

$$\left(\int_{-\infty}^{\infty} e^{-x^2}\, dx\right)^2 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-(x^2+y^2)}\, dxdy$$

$$= \int_{0}^{\infty} \int_{-\pi}^{\pi} e^{-r^2}\, rdrd\theta = 2\pi \int_{0}^{\infty} e^{-r^2} rdr = \pi,$$

which yields

$$\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi} \tag{7.1.15}$$

after taking the square-root. Hence, for any $\alpha > 0$,

$$\int_{-\infty}^{\infty} e^{-\alpha x^2} dx = \sqrt{\frac{\pi}{\alpha}}, \tag{7.1.16}$$

which implies that (7.1.14) holds, by choosing $\alpha = 1/(4\sigma^2)$.

To compute the Fourier transform of $g_\alpha(x)$, we first consider $v(x) = e^{-x^2}$ and formulate its Fourier transform as

$$G(\omega) = \widehat{v}(\omega) = \int_{-\infty}^{\infty} e^{-x^2} e^{-i\omega x} dx$$

$$= \int_{-\infty}^{\infty} e^{-(x^2+i\omega x)} dx.$$

Then for $y \in \mathbb{R}$, we have

$$G(-iy) = \int_{-\infty}^{\infty} e^{-(x^2+yx)} dx$$

$$= e^{y^2/4} \int_{-\infty}^{\infty} e^{-(x+y/2)^2} dx = \sqrt{\pi} e^{y^2/4}. \tag{7.1.17}$$

Hence, when the function

$$H(z) = G(z) - \sqrt{\pi} e^{-z^2/4} \tag{7.1.18}$$

is considered as a function of a complex variable z, $H(z)$ is analytic for all $z \in \mathbb{C}$ (or $H(z)$ is called an entire function) and $H(-iy) = 0$ for all $y \in \mathbb{R}$. Recall that if an analytic function vanishes on a set with a finite number of accumulation points, then the function vanishes identically in the domain of analyticity. Hence, $H(z) = 0$ for all $z \in \mathbb{C}$, so that

$$G(\omega) - \sqrt{\pi} e^{-\omega^2/4} = 0, \ \omega \in \mathbb{R};$$

or

$$\widehat{v}(\omega) = \int_{-\infty}^{\infty} e^{-x^2} e^{-i\omega x} dx = \sqrt{\pi} e^{-\omega^2/4}. \tag{7.1.19}$$

This enables us to compute the Fourier transform $\widehat{g}_\sigma(\omega)$ of $g_\sigma(x)$; namely, in view of (7.1.13) and (7.1.19), we have

$$
\begin{aligned}
\widehat{g}_\sigma(\omega) &= \frac{1}{2\sigma\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-\left(\frac{x}{2\sigma}\right)^2} e^{-i\omega x} dx \\
&= \frac{1}{2\sigma\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-y^2} e^{-i(2\sigma\omega)y} (2\sigma) dy \\
&= \frac{1}{\sqrt{\pi}} \sqrt{\pi}\, e^{-(2\sigma\omega)^2/4} = e^{-(\sigma\omega)^2}.
\end{aligned}
$$

Theorem 3 | **Gaussian is preserved under Fourier transform** | *The Fourier transform of the Gaussian function $g_\sigma(x)$, $\sigma > 0$, defined in (7.1.13), is*

$$\widehat{g}_\sigma(\omega) = e^{-\sigma^2\omega^2}. \tag{7.1.20}$$

Remark 4 By (7.1.19) and the change of variables of integration, we have

$$\int_{-\infty}^{\infty} e^{-\sigma^2\omega^2} e^{ix\omega}\, d\omega = \frac{\sqrt{\pi}}{\sigma} e^{-\left(\frac{x}{2\sigma}\right)^2} = 2\pi g_\sigma(x). \tag{7.1.21}$$

The detailed verification is left as an exercise (see Exercise 21).   ∎

By allowing $\sigma > 0$ in $g_\sigma(x)$ to be a free parameter, we have a powerful mathematical tool, called "positive approximate identity" $\{g_\sigma\}$ for $L_2(\mathbb{R})$ (see the family $\{\sigma_n\}$ of Fejér's kernels in Theorem 3 on p.293 as a positive approximate identity for $PC_{2\pi}^\star$, which obviously extends to all periodic functions in $L_2[a, b]$, $-\infty < a < b < \infty$).

Theorem 4 | **Positive approximate identity** | *Let $g_\sigma(x)$ be the Gaussian function in (7.1.13). Then*

(a) $g_\sigma(x) \geq 0$ *for all $x \in \mathbb{R}$;*
(b) $\int_{-\infty}^{\infty} g_\sigma(x) dx = 1$, *all $\sigma > 0$; and*
(c) *for any $\delta > 0$,*

$$\lim_{\sigma \to 0} \int_{|x| > \delta} g_\sigma(x) dx = 0.$$

Clearly $g_\sigma(x) > 0$ for any $x \in \mathbb{R}$, while the property (b) has been proved above. The proof of the result (c) is left as an exercise (see Exercise 20).

The property of positive approximate identity of $g_\sigma(x)$ is applied in the following to show that convolution of $g_\sigma(x)$ preserves all bounded continuous functions, as $\sigma \to 0^+$. In other words, the limit of $g_\sigma(x)$ is the "delta function".

Theorem 5 *Let $f \in L_\infty(\mathbb{R})$ be continuous at $x = x_0$. Then*

$$\lim_{\sigma \to 0+} (f * g_\sigma)(x_0) = f(x_0).$$

Proof Let $\epsilon > 0$ be arbitrarily given. Since there exists some $\delta > 0$, such that

$$|f(x) - f(x_0)| < \frac{\epsilon}{2}$$

for $|x - x_0| \leq \delta$, it follows, by applying (a) and (b) in Theorem 4, that

$$
\begin{aligned}
|f(x_0) - (f * g_\sigma)(x_0)| &= \left| \int_{-\infty}^{\infty} \big(f(x_0) - f(x_0 - t)\big) g_\sigma(t) dt \right| \\
&\leq \int_{-\infty}^{\infty} |f(x_0) - f(x_0 - t)| g_\sigma(t) dt \\
&= \int_{-\delta}^{\delta} |f(x_0) - f(x_0 - t)| g_\sigma(t) dt \\
&\quad + \int_{|t|>\delta} |f(x_0) - f(x_0 - t)| g_\sigma(t) dt \\
&\leq \max_{|x-x_0|\leq\delta} |f(x_0) - f(x)| \int_{-\delta}^{\delta} g_\sigma(t) dt \\
&\quad + 2\|f\|_\infty \int_{|t|>\delta} g_\sigma(t) dt \\
&\leq \frac{\epsilon}{2} + 2\|f\|_\infty \int_{|t|>\delta} g_\sigma(t) dt.
\end{aligned}
\qquad (7.1.22)
$$

Hence, since $f \in L_\infty(\mathbb{R})$, we may apply (c) of Theorem 4 to deduce the existence of some $\sigma_0 > 0$, such that

$$\int_{|t|>\delta} g_\sigma(t) dt \leq \frac{\epsilon}{4\|f\|_\infty}$$

for all $0 < \sigma < \sigma_0$. This completes the proof of Theorem 5.   ∎

Exercises

Exercise 1 Let

$$h_1(x) = \begin{cases} 1, & \text{if } 0 \leq x \leq 1, \\ 0, & \text{otherwise.} \end{cases}$$

Compute the Fourier transform $\widehat{h}_1(\omega)$ of $h_1(x)$.

Exercise 2 Show that if $f \in L_1(\mathbb{R})$, then $\widehat{f} = \mathbb{F}f \in L_\infty(\mathbb{R})$ and $\|\widehat{f}\|_\infty \leq \|f\|_1$.

Exercise 3 Show that if $f \in L_1(\mathbb{R})$, $f'(x)$ exists for all $x \in \mathbb{R}$ and $f' \in L_1(\mathbb{R})$, then $(\mathbb{F}f')(\omega) = i\omega\widehat{f}(\omega)$ by applying integration by parts.

Exercise 4 If $f, g \in L_2(\mathbb{R})$, show that $f * g \in L_\infty(\mathbb{R})$.

Exercise 5 If $f, g \in L_1(\mathbb{R})$, show that $f * g \in L_1(\mathbb{R})$.

Exercise 6 Let $h_2(x)$ be defined by $h_2(x) = (h_1 * h_1)(x)$, where $*$ is the convolution operation and $h_1(x)$ is defined in Exercise 1.

(a) Show that $h_2(x) = 0$ for $x < 0$ or $x > 2$.
(b) Compute $h_2(x)$ for $0 \leq x \leq 1$.
(c) Compute $h_2(x)$ for $1 \leq x \leq 2$.
(d) Graph $y = h_2(x)$.

Exercise 7 By applying the result in Exercise 1 and formula (7.1.12) of Theorem 2, compute the Fourier transform $\widehat{h}_2(\omega)$ of $h_2(x)$.

Exercise 8 By applying (7.1.9) of Theorem 2 and the result in Exercise 7, compute the Fourier transform $\widehat{H}_2(\omega)$ of the function $H_2(x) = h_2(x + 1)$.

Exercise 9 Extend the formula (7.1.12) of Theorem 2 to write out the Fourier transform of $g(x) = f_1 * \big(f_2 * (f_3 * f_4)\big)(x)$, where $f_j \in L_1(\mathbb{R})$, $1 \leq j \leq 4$.

Exercise 10 In Exercise 9, show that the Fourier transform $\widehat{g}(\omega)$ of $g(x)$ is independent of the permutation of f_1, f_2, f_3, f_4 in the convolution formulation, and hence conclude that $g(x)$ may be written as $g(x) = (f_1 * f_2 * f_3 * f_4)(x)$.

Exercise 11 Verify the result in Exercise 10 directly from the definition of convolution without appealing to Fourier transform.

Exercise 12 As a result of Exercise 10, let

$$h_m(x) = \underbrace{(h_1 * h_1 * \cdots * h_1)}_{m \text{ copies of } h_1}(x),$$

where $h_1(x)$ is defined in Exercise 1. Computer the Fourier transform $\widehat{h}_m(\omega)$.

Exercise 13 For the function $h_m(x)$ in Exercise 12, define $H_m(x) = h_m(x + \frac{m}{2})$. Compute $\widehat{H}_m(\omega)$ and show that $\widehat{H}_m(\omega)$ is a real-valued function.

Exercise 14 Show that for $t \in \mathbb{R}$,

$$|e^{-it} - 1| \leq |t|.$$

Exercise 15 Derive the formula (7.1.8).

Exercise 16 Derive the formula (7.1.9).

Exercise 17 Derive the formula (7.1.10).

Exercise 18 Derive the formula (7.1.11).

Exercise 19 Derive the formulas (7.1.16) and (7.1.14) by applying (7.1.15).

Exercise 20 Prove the property (c) of Theorem 4.

Exercise 21 Derive the formula (7.1.21).

7.2 Inverse Fourier Transform and Sampling Theorem

The definition of the Fourier transform for $L_1(\mathbb{R})$ in (7.1.1) on p.320 can be extended to functions $f \in L_2(\mathbb{R})$ by considering the truncated functions

$$f_N(x) = \begin{cases} f(x), & \text{for} |x| \leq N, \\ 0, & \text{for} |x| > N, \end{cases} \tag{7.2.1}$$

for $N = 1, 2, \cdots$. Observe that each f_N is compactly supported. Thus, from the assumption $f \in L_2(\mathbb{R})$, we have $f_N \in (L_2 \cap L_1)(\mathbb{R})$, so that $\widehat{f_N}$ is well-defined. In addition, since $\{f_N\}$ converges to f in $L_2(\mathbb{R})$, $\{f_N\}$ is a Cauchy sequence in $L_2(\mathbb{R})$. In this section, by applying the Gaussian function and its Fourier transform, we will prove that for $N = 1, 2, \cdots$,

$$\|\widehat{f_N}\|_2^2 = 2\pi \|f_N\|_2^2 \tag{7.2.2}$$

(see (7.2.4)–(7.2.6)). It then follows from (7.2.2) and the fact that $\{f_N\}$ is a Cauchy sequence in $L_2(\mathbb{R})$, that the sequence $\{\widehat{f_N}(\omega)\}_N$ is also a Cauchy sequence in $L_2(\mathbb{R})$, and its limit, being a function in $L_2(\mathbb{R})$, can be used as the definition of the Fourier transform of f.

Definition 1 $\boxed{\textbf{Fourier transforms of } L_2(\mathbb{R}) \textbf{ functions}}$ *Let $f \in L_2(\mathbb{R})$ and $f_N(x)$ be the truncations of f defined by (7.2.1). Then the Fourier transform $\widehat{f}(\omega)$ of f is defined as the limit of $\{\widehat{f_N}(\omega)\}_N$ in $L_2(\mathbb{R})$.*

Remark 1 The Fourier transform $\widehat{f}$ for $f \in L_2(\mathbb{R})$ defined in Definition 1 is independent of the choice of $\{f_N\}$ in the sense that if $\{g_N\}$ with $g_N \in (L_2 \cap L_1)(\mathbb{R})$ converges to f in $L_2(\mathbb{R})$, then $\{\widehat{g_N}\}$ has the same limit as $\{\widehat{f_N}\}$. Indeed, by (7.2.2), we have

$$\|\widehat{g_N} - \widehat{f_N}\|_2 = 2\pi \|g_N - f_N\|_2$$
$$\leq 2\pi \|g_N - f\|_2 + 2\pi \|f - f_N\|_2 \to 0$$

as $N \to \infty$. Thus, $\{\widehat{g}_N\}$ and $\{\widehat{f}_N\}$ have the same limit. ∎

Recall from (6.1.10) of Theorem 2 on p.271 that the "energy" of any $f \in PC^{\star}_{2\pi}$ is preserved by the energy of the sequence $\{c_k\}$ of Fourier coefficients of f, called Parseval's identity for $PC^{\star}_{2\pi}$. The analogue of this result extends to the Fourier transform via the identity (7.2.2) to be established in the following theorem.

Theorem 1 | **Parseval's theorem** | *Let $\widehat{f}(\omega)$ be the Fourier transform of $f \in L_2(\mathbb{R})$. Then*

$$\|\widehat{f}\|_2^2 = 2\pi \|f\|_2^2. \tag{7.2.3}$$

In other words, the "energy" of $f(x)$ is "preserved" by its Fourier transform $\widehat{f}(\omega)$. The identity (7.2.3) is commonly called Plancherel's formula but will be called Parseval's identity in this book for uniformity.

Proof of Theorem 1 | By the definition of the Fourier transform for functions in $L_2(\mathbb{R})$, it is sufficient to derive the identity (7.2.3) for the truncated functions; and in view of Remark 1, we may assume that $f \in (L_1 \cap L_2)(\mathbb{R})$ and consider its corresponding "autocorrelation function" $F(x)$, defined by

$$F(x) = \int_{-\infty}^{\infty} f(t)\overline{f(t-x)}dt. \tag{7.2.4}$$

By setting

$$f^-(x) = f(-x),$$

the autocorrelation can be viewed as the convolution of f and $\overline{f^-}$, namely:

$$F(x) = (f * \overline{f^-})(x),$$

so that it follows from Remark 1 on p.322 that $F(x)$ is in both L_∞ and $L_1(\mathbb{R})$. Furthermore, by (v) in Theorem 2 on p.323, we have

$$\widehat{F}(\omega) = \widehat{f}(\omega)\overline{\widehat{f}(\omega)} = |\widehat{f}(\omega)|^2 \tag{7.2.5}$$

(see Exercise 1). Therefore, by applying (7.2.5) together with (7.1.20) on p.326, we obtain

$$\int_{-\infty}^{\infty} |\widehat{f}(\omega)|^2 e^{-\sigma^2\omega^2}d\omega = \int_{-\infty}^{\infty} \widehat{f}(\omega)\overline{\widehat{f}(\omega)}\,\widehat{g}_\sigma(\omega)d\omega$$

$$= \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{\infty} f(y)e^{-i\omega y}\int_{-\infty}^{\infty} \overline{f(x)}e^{i\omega x}\widehat{g}_\sigma(\omega)dxdy \right\}d\omega$$

$$= \int_{-\infty}^{\infty} f(y)\int_{-\infty}^{\infty} \overline{f(x)}\left\{ \int_{-\infty}^{\infty} \widehat{g}_\sigma(\omega)e^{i(x-y)\omega}d\omega \right\}dxdy$$

$$= 2\pi \int_{-\infty}^{\infty} f(y) \int_{-\infty}^{\infty} \overline{f(x)} g_\sigma(x-y) dx dy$$

$$= 2\pi \int_{-\infty}^{\infty} f(y) \int_{-\infty}^{\infty} \overline{f(y+t)} g_\sigma(t) dt dy$$

$$= 2\pi \int_{-\infty}^{\infty} F(-t) g_\sigma(t) dt = 2\pi (F * g_\sigma)(0),$$

where the 4th equality follows from (7.1.21) on p.326 and the change of variables of integration $t = x - y$ is applied to derive the second last line. In addition, since $F(x)$ is a continuous function (see Exercise 2), it follows from Theorem 5 on p.327 with $x_0 = 0$ that by taking the limit as $\sigma \to 0$,

$$\|\widehat{f}\|_2^2 = 2\pi F(0) = 2\pi \|f\|_2^2. \tag{7.2.6}$$

This completes the proof of Theorem 1. ∎

As a consequence of Theorem 1, we have the following result.

Theorem 2 | **Parseval's formula** | *Let $f, g \in L_2(\mathbb{R})$. Then*

$$\langle f, g \rangle = \frac{1}{2\pi} \langle \widehat{f}, \widehat{g} \rangle. \tag{7.2.7}$$

We only derive (7.2.7) for real-valued functions and leave the complex-valued setting as an exercise (see Exercise 3). For real-valued $f(x)$ and $g(x)$, since

$$\|f \pm g\|_2^2 = \|f\|_2^2 \pm 2\langle f, g \rangle + \|g\|_2^2,$$

we have

$$\langle f, g \rangle = \frac{1}{4} \Big(\|f + g\|_2^2 - \|f - g\|_2^2 \Big).$$

Hence, it follows from (7.2.3) (with f replaced by $(f+g)$ and $(f-g)$, respectively) that

$$\langle f, g \rangle = \frac{1}{4} \frac{1}{2\pi} \Big(\|\widehat{f} + \widehat{g}\|_2^2 - \|\widehat{f} - \widehat{g}\|_2^2 \Big)$$

$$= \frac{1}{2\pi} \langle \widehat{f}, \widehat{g} \rangle.$$

∎

The following result, though similar to the formulation of (7.2.7), is more elementary and can be proved by a straightforward change of order of integrations (that is, by applying Fubini's theorem).

Theorem 3 *Let $f, g \in L_2(\mathbb{R})$. Then*

$$\langle \widehat{f}, \overline{g} \rangle = \langle \widehat{g}, \overline{f} \rangle. \tag{7.2.8}$$

To prove this theorem, it is sufficient to derive (7.2.8) for $f, g \in (L_1 \cap L_2)(\mathbb{R})$, since the completion of $(L_1 \cap L_2)(\mathbb{R})$ in $L_2(\mathbb{R})$ is $L_2(\mathbb{R})$. Under this additional assumption, we may apply Fubini's theorem to interchange the order of integrations. Thus, we have

$$\langle \widehat{f}, \overline{g} \rangle = \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{\infty} f(x) e^{-iyx} \, dx \right\} g(y) dy$$

$$= \int_{-\infty}^{\infty} f(x) \left\{ \int_{-\infty}^{\infty} g(y) e^{-ixy} \, dy \right\} dx$$

$$= \int_{-\infty}^{\infty} f(x) \widehat{g}(x) dx = \langle \widehat{g}, \overline{f} \rangle.$$

∎

To be able to recover $f(x)$ from its Fourier transform $\widehat{f}(\omega) = (\mathbb{F}f)(\omega)$, we need to introduce the inverse $\mathbb{F}^{-1}$ of the Fourier transform operation $\mathbb{F}$. To accomplish this goal, let us first consider the companion transform $\mathbb{F}^{\#}$ of $\mathbb{F}$, defined by

$$(\mathbb{F}^{\#} g)(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} g(\omega) e^{ix\omega} d\omega \qquad (7.2.9)$$

for $g \in L_1(\mathbb{R})$.

Remark 2 Observe that $\mathbb{F}^{\#}$ introduced in (7.2.9) is related to the Fourier transform $\mathbb{F}$ in (7.1.1) on p.320 by

$$(\mathbb{F}^{\#} g)(x) = \frac{1}{2\pi} (\mathbb{F}g)(-x) = \frac{1}{2\pi} \overline{(\mathbb{F}\overline{g})(x)}. \qquad (7.2.10)$$

∎

Under the additional assumption that $\widehat{f} \in L_1(\mathbb{R})$, we can recover $f \in L_2(\mathbb{R})$ from $\widehat{f}(\omega)$ by applying $\mathbb{F}^{\#}$, as in the following theorem.

Theorem 4 $\boxed{\text{Inverse Fourier transform}}$ *Let $\widehat{f}(\omega)$ be the Fourier transform of $f \in L_2(\mathbb{R})$. Suppose that $\widehat{f} \in L_1(\mathbb{R})$. Then*

$$f(x) = \left(\mathbb{F}^{\#} \widehat{f} \right)(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \widehat{f}(\omega) e^{ix\omega} \, d\omega. \qquad (7.2.11)$$

That is, $\mathbb{F}^{\#} = \mathbb{F}^{-1}$, to be called the inverse Fourier transform (or IFT).

Remark 3 There exist functions $f \in L_p(\mathbb{R})$ for all p, $1 \leq p \leq \infty$, such that $\widehat{f} \notin L_1(\mathbb{R})$, as illustrated in the following example. Hence, $\mathbb{F}^{\#} = \mathbb{F}^{-1}$ in Theorem 4 is restricted only to operation on functions $\widehat{f} \in L_1(\mathbb{R})$. ∎

Example 1 *Let $f_0(x)$ be defined by*

$$f_0(x) = \begin{cases} e^{-x}, & \text{for } x \geq 0, \\ 0, & \text{for } x < 0, \end{cases}$$

and $\widehat{f_0}(\omega)$ be its Fourier transform. Then $f_0 \in L_p(\mathbb{R})$ for all p, $1 \leq p \leq \infty$, but $\widehat{f_0} \notin L_1(\mathbb{R})$.

Solution It is clear that $f_0 \in L_p(\mathbb{R})$, since

$$\int_{-\infty}^{\infty} |f_0(x)|^p dx = \int_0^{\infty} e^{-px} dx = \frac{1}{p} < \infty,$$

for $1 \leq p < \infty$ and $\|f_0\|_{\infty} = 1$. On the other hand,

$$\begin{aligned} \widehat{f_0}(\omega) &= \int_0^{\infty} e^{-x} e^{-ix\omega} dx = \int_0^{\infty} e^{-x(1+i\omega)} dx \\ &= \frac{1}{1+i\omega}, \end{aligned}$$

which is not in $L_1(\mathbb{R})$. Indeed,

$$\begin{aligned} \int_{-\infty}^{\infty} |\widehat{f_0}(\omega)| d\omega &= \int_{-\infty}^{\infty} (1+\omega^2)^{-1/2} d\omega \\ &= 2 \int_0^{\infty} (1+\omega^2)^{-1/2} d\omega > 2 \int_1^{\infty} (1+\omega^2)^{-1/2} d\omega \\ &> \sqrt{2} \int_1^{\infty} (\omega^2)^{-1/2} d\omega = \sqrt{2} \ln \omega \Big|_{\omega=1}^{\infty} = \infty. \quad \blacksquare \end{aligned}$$

Proof of Theorem 4 To prove Theorem 4, we set $g = \widehat{f}$ and apply (7.2.10), Theorem 3, and Theorem 2, consecutively, to compute

$$\begin{aligned} \|f - \mathbb{F}^{\#} g\|_2^2 &= \|f\|_2^2 - \langle f, \mathbb{F}^{\#} g \rangle - \langle \mathbb{F}^{\#} g, f \rangle + \|\mathbb{F}^{\#} g\|_2^2 \\ &= \|f\|_2^2 - \frac{1}{2\pi} \langle f, \overline{\mathbb{F} g} \rangle - \frac{1}{2\pi} \langle \overline{\mathbb{F} g}, f \rangle + (\frac{1}{2\pi})^2 \|\mathbb{F} g\|_2^2 \\ &= \|f\|_2^2 - \frac{1}{2\pi} \langle \mathbb{F} g, \overline{f} \rangle - \frac{1}{2\pi} \overline{\langle \mathbb{F} g, \overline{f} \rangle} + (\frac{1}{2\pi})^2 (2\pi) \|\overline{g}\|_2^2 \\ &= \|f\|_2^2 - \frac{1}{2\pi} \langle \widehat{f}, g \rangle - \frac{1}{2\pi} \overline{\langle \widehat{f}, g \rangle} + \frac{1}{2\pi} \|g\|_2^2 \\ &= \|f\|_2^2 - \frac{1}{2\pi} \langle \widehat{f}, \widehat{f} \rangle - \frac{1}{2\pi} \overline{\langle \widehat{f}, \widehat{f} \rangle} + \frac{1}{2\pi} \|\widehat{f}\|_2^2 \\ &= \|f\|_2^2 - \frac{1}{2\pi} \|\widehat{f}\|_2^2 - \frac{1}{2\pi} \|\widehat{f}\|_2^2 + \frac{1}{2\pi} \|\widehat{f}\|_2^2 \\ &= \|f\|_2^2 - \frac{1}{2\pi} \|\widehat{f}\|_2^2 = \|f\|_2^2 - \|f\|_2^2 = 0. \end{aligned}$$

Here, we have used the fact that $\|\bar{h}\|_2 = \|h\|_2$ and $g = \widehat{f}$. Hence,

$$(f - \mathbb{F}^{\#}g)(x) = 0$$

for almost all x, which implies the validity of (7.2.11). ∎

Next, let us recall the concept of "ideal lowpass filtering" introduced in Remark 2 on p.289 for functions in $PC_{2\pi}^{\star}$. By considering Dirichlet's kernels $D_n(x)$ defined in (6.3.2) on p.288 as the "convolution filters" of the signal $f(x)$, the output is the nth-order partial sum $(S_n f)$ of the Fourier series Sf. In other words, this ideal lowpass filtering process retains the low-frequency coefficients $c_k(f)$ with $|k| \le n$ but removes the high-frequency content $c_k(f)$ for all k with $|k| > n$ from Sf. Hence, the precise analogy of ideal lowpass filtering of signals $f \in L_2(\mathbb{R})$ is removing the frequency content $\widehat{f}(\omega)$ of $f(x)$ for all ω with $|\omega| > \eta$, where $\eta > 0$ is any desired "cut-off" frequency. More precisely, let

$$\chi_{[-\eta,\eta]}(\omega) = \begin{cases} 1, & \text{for} |\omega| \le \eta, \\ 0, & \text{for} |\omega| > \eta, \end{cases} \tag{7.2.12}$$

denote the characteristic function of the interval $[-\eta, \eta]$. Then ideal lowpass filtering of $f(x)$ is mapping

$$\widehat{f}(\omega) \longrightarrow \widehat{f}(\omega)\chi_{\eta}(\omega).$$

Therefore, in view of the result in (7.1.12) of Theorem 2, it is clear that the ideal lowpass filter is the function $h_{\eta}(x)$ with Fourier transform given by

$$\widehat{h}_{\eta}(\omega) = \chi_{[-\eta,\eta]}(\omega). \tag{7.2.13}$$

Hence, since $\chi_{\eta} \in (L_1 \cap L_2)(\mathbb{R})$, it follows from Theorem 4 that

$$\begin{aligned} h_{\eta}(x) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \widehat{h}_{\eta}(\omega)e^{ix\omega}\, d\omega \\ &= \frac{1}{2\pi} \int_{-\eta}^{\eta} e^{ix\omega}\, d\omega = \frac{1}{2\pi}\frac{e^{i\eta x} - e^{-i\eta x}}{ix} \\ &= \frac{\sin \eta x}{\pi x} = \frac{\eta}{\pi}\left(\frac{\sin \eta x}{\eta x}\right). \end{aligned} \tag{7.2.14}$$

Definition 2 ┃ **Sinc function** ┃ *The* "sinc" *function is defined by*

$$\mathrm{sinc}(x) = \frac{\sin \pi x}{\pi x}. \tag{7.2.15}$$

Remark 4 Let $\eta > 0$. Then by applying (7.2.14), the ideal lowpass filter $h_{\eta}(x)$ can be formulated as

$$h_\eta(x) = \frac{\eta}{\pi} \text{sinc}\left(\frac{\eta}{\pi} x\right). \tag{7.2.16}$$

Thus by (7.2.13),

$$\left(\text{sinc}(\frac{\eta}{\pi} x)\right)^\wedge (\omega) = \frac{\pi}{\eta} \widehat{h}_\eta(\omega) = \frac{\pi}{\eta} \chi_{[-\eta,\eta]}(\omega). \tag{7.2.17}$$

It follows from (7.1.12) of Theorem 2 on p. 323 that the convolution operation

$$(f * h_\eta)(x) = \int_{-\infty}^\infty f(t) h_\eta(x-t) dt$$

applied to any $f \in L_2(\mathbb{R})$ is an ideal lowpass filtering of $f(x)$, considered as a signal, in the sense that

$$(f * h_\eta)^\wedge (\omega) = \widehat{f}(\omega) \chi_{[-\eta,\eta]}(\omega).$$

∎

We end this section by applying Theorem 4 to derive the following so-called "sampling theorem" which assures perfect reconstruction of band-limited continuous functions (considered as analog signals) from certain discrete samples.

Definition 3 $\boxed{\text{Band-limited function}}$ *A function $f \in L_2(\mathbb{R})$ is said to be band-limited if its Fourier transform $\widehat{f}(\omega)$ vanishes for all ω with $|\omega| > \sigma$ for some $\sigma > 0$. The bandwidth of a band-limited function is 2σ, where $\sigma > 0$ is the smallest value for which $\widehat{f}(\omega) = 0$ for $|\omega| > \sigma$.*

Let f be a band-limited function with $\widehat{f}(\omega) = 0$ for $|\omega| > \sigma$. Since $f \in L_2(\mathbb{R})$, we have $\widehat{f} \in L_2(\mathbb{R})$. In addition,

$$\int_\mathbb{R} |\widehat{f}(\omega)| d\omega = \int_{-\sigma}^\sigma |\widehat{f}(\omega)| d\omega \leq \sqrt{2\sigma} \left(\int_{-\sigma}^\sigma |\widehat{f}(\omega)|^2 d\omega\right)^{1/2} < \infty.$$

That is, $\widehat{f} \in L_1(\mathbb{R})$. Thus it follows from Theorem 4 that

$$f(x) = \frac{1}{2\pi} \int_{-\sigma}^\sigma \widehat{f}(\omega) e^{ix\omega} d\omega. \tag{7.2.18}$$

By (ii) in Theorem 1 on p.320, we observe that the function defined by the integral on the right-hand side of (7.2.18) is uniformly continuous on $\mathbb{R}$, since $\widehat{f} \in L_1(\mathbb{R})$. Thus by re-defining the original f at certain points, if necessary, f is continuous on $\mathbb{R}$. In the following, a band-limited function is a continuous function with representation (7.2.18) for some function $g(\omega)$ on $(-\sigma, \sigma)$ (which happens to be $\widehat{f}(\omega)$). In fact, since for any $k \geq 1$, $\omega^k \widehat{f}(\omega) \in (L_2 \cap L_1)(\mathbb{R})$, by (v) in Theorem

1 on p.320, we see that f given by (7.2.18) is in $C^k(\mathbb{R})$. In summary, we have the following theorem.

Theorem 5 *A band-limited function is in $C^\infty(\mathbb{R})$.*

Next we introduce the sampling theorem, which is also called the Nyquist-Shannon Sampling Theorem.

Theorem 6 $\boxed{\textbf{Sampling theorem}}$ *Let $f \in L_2(\mathbb{R})$ be band-limited with bandwidth not exceeding 2σ. Then $f(x)$ can be recovered from the discrete samples $f(\frac{k\pi}{\sigma})$, $k \in \mathbb{Z}$, by applying the formula*

$$f(x) = \sum_{k=-\infty}^{\infty} f\left(\frac{k\pi}{\sigma}\right) \text{sinc}\left(\frac{\sigma}{\pi}x - k\right), \quad x \in \mathbb{R}, \tag{7.2.19}$$

where the series converges to $f(x)$ both uniformly in $\mathbb{R}$ and in $L_2(\mathbb{R})$.

Proof Since $f \in L_2(\mathbb{R})$, we have $\widehat{f} \in L_2[-\sigma, \sigma]$. Thus, by Theorem 3 on p.271, $\widehat{f}$ can be represented by its Fourier series

$$\widehat{f}(\omega) = \sum_{k=-\infty}^{\infty} c_k(\widehat{f}) e^{ik\pi\omega/\sigma}, \quad \omega \in [-\pi, \pi],$$

which converges to $\widehat{f}(\omega)$, in the sense that the sequence of the partial sums

$$(S_n \widehat{f})(\omega) = \sum_{k=-n}^{n} c_k(\widehat{f}) e^{ik\pi\omega/\sigma}$$

of the Fourier series converges to $\widehat{f}(\omega)$ in $L_2[-\sigma, \sigma]$. Here, $c_k(\widehat{f})$, $k \in \mathbb{Z}$, are the Fourier coefficients of $\widehat{f}(\omega)$ given by

$$\begin{aligned} c_k(\widehat{f}) &= \frac{1}{2\sigma} \int_{-\sigma}^{\sigma} \widehat{f}(\omega) \, e^{-ik\pi\omega/\sigma} \, d\omega \\ &= \frac{\pi}{\sigma} f\left(\frac{-k\pi}{\sigma}\right), \end{aligned}$$

where (7.2.18) is applied to yield the last equality.

Let $L_n(x)$ be the nth partial sum of the series in (7.2.19):

$$L_n(x) = \sum_{k=-n}^{n} f\left(\frac{k\pi}{\sigma}\right) \text{sinc}\left(\frac{\sigma}{\pi}x - k\right).$$

Observe that

$$\text{sinc}\left(\frac{\sigma}{\pi}x - k\right) = \frac{\sin\left(\sigma(x - \frac{k\pi}{\sigma})\right)}{\sigma(x - \frac{k\pi}{\sigma})} = \frac{1}{2\sigma}\int_{-\sigma}^{\sigma}e^{i(x-\frac{k\pi}{\sigma})\omega}d\omega.$$

Thus,

$$
\begin{aligned}
L_n(x) &= \frac{1}{2\sigma}\sum_{k=-n}^{n}f\left(\frac{k\pi}{\sigma}\right)\int_{-\sigma}^{\sigma}e^{i(x-\frac{k\pi}{\sigma})\omega}d\omega \\
&= \frac{1}{2\sigma}\int_{-\sigma}^{\sigma}\sum_{k=-n}^{n}f\left(\frac{k\pi}{\sigma}\right)e^{-ik\pi\omega/\sigma}e^{ix\omega}d\omega \\
&= \frac{1}{2\pi}\int_{-\sigma}^{\sigma}\frac{\pi}{\sigma}\left\{\sum_{j=-n}^{n}f\left(\frac{-j\pi}{\sigma}\right)e^{ij\pi\omega/\sigma}\right\}e^{ix\omega}d\omega \\
&= \frac{1}{2\pi}\int_{-\sigma}^{\sigma}(S_n\widehat{f})(\omega)e^{ix\omega}d\omega.
\end{aligned}
$$

This, together with (7.2.18), leads to

$$L_n(x) - f(x) = \frac{1}{2\pi}\int_{-\sigma}^{\sigma}\left((S_n\widehat{f})(\omega) - \widehat{f}(\omega)\right)e^{ix\omega}d\omega.$$

Therefore,

$$
\begin{aligned}
|L_n(x) - f(x)| &\leq \frac{1}{2\pi}\int_{-\sigma}^{\sigma}\left|(S_n\widehat{f})(\omega) - \widehat{f}(\omega)\right|\left|e^{ix\omega}\right|d\omega \\
&\leq \frac{1}{2\pi}\sqrt{2\sigma}\left(\int_{-\sigma}^{\sigma}\left|(S_n\widehat{f})(\omega) - \widehat{f}(\omega)\right|^2 d\omega\right)^{\frac{1}{2}} \to 0,
\end{aligned}
$$

as $n \to \infty$ for all $x \in \mathbb{R}$, since $(S_n\widehat{f})(\omega)$ converges to $\widehat{f}(\omega)$ in $L_2[-\sigma, \sigma]$. This shows that the series in (7.2.19) converges to $f(x)$ uniformly in $\mathbb{R}$.

Next we show that $L_n(x)$ converges to $f(x)$ in $L_2(\mathbb{R})$. By (7.2.17), we have (see Exercise 7)

$$\left(\text{sinc}(\frac{\sigma}{\pi}x - k)\right)^{\wedge}(\omega) = \frac{\pi}{\sigma}\chi_{[-\sigma,\sigma]}(\omega)e^{-ik\pi\omega/\sigma}. \tag{7.2.20}$$

Thus,

$$\widehat{L}_n(\omega) = \frac{\pi}{\sigma}\sum_{k=-n}^{n}f\left(\frac{k\pi}{\sigma}\right)\chi_{[-\sigma,\sigma]}(\omega)e^{-ik\pi\omega/\sigma} = (S_n\widehat{f})(\omega)\chi_{[-\sigma,\sigma]}(\omega).$$

Hence, it follows from Parseval's theorem on p.330 that

$$2\pi \|L_n - f\|_2^2 = \|\widehat{L}_n - \widehat{f}\|_2^2$$
$$= \int_{-\infty}^{\infty} \left| (S_n \widehat{f})(\omega)\chi_{[-\sigma,\sigma]}(\omega) - \widehat{f}(\omega) \right|^2 d\omega$$
$$= \int_{-\sigma}^{\sigma} \left| (S_n \widehat{f})(\omega) - \widehat{f}(\omega) \right|^2 d\omega \to 0,$$

as $n \to \infty$, since $(S_n \widehat{f})(\omega)$ converges to $\widehat{f}(\omega)$ in $L_2[-\sigma, \sigma]$. That is, the series in (7.2.19) converges to $f(x)$ in $L_2[-\sigma, \sigma]$.   ∎

Exercises

Exercise 1 Show that the Fourier transform of the autocorrelation function $F(x)$ of $f(x)$ is the square of the magnitude of the Fourier transform of $f(x)$ as in (7.2.5).

Exercise 2 Prove that the autocorrelation function $F(x)$ defined in (7.2.4) is a continuous function for any function $f \in L_2(\mathbb{R})$.

Exercise 3 Derive (7.2.7) by applying (7.2.3) for complex-valued functions $f, g \in L_2(\mathbb{R})$.
   Hint: In addition to considering $\|f \pm g\|_2^2$ as in the proof for the real-valued setting, also consider $\|f \pm ig\|_2^2$ to obtain $\text{Im}\langle \widehat{f}, g \rangle$.

Exercise 4 For $y \in \mathbb{R}$, let $y_+ = \max(y, 0)$ and consider the function $g(\omega) = (1 - |\omega|)_+$. Show that $(\mathbb{F}^{\#} g)(x) = f(x)$, where

$$f(x) = \begin{cases} \frac{1 - \cos x}{\pi x^2}, & \text{if } x \neq 0, \\ \frac{1}{2\pi}, & \text{if } x = 0. \end{cases}$$

Hence, apply Theorem 4 to conclude that $\widehat{f}(\omega) = (1 - |\omega|)_+$.

Exercise 5 Show that the function $f(x)$ in Exercise 4 is band-limited with bandwidth equal to $2\sigma = 2$ (or $\sigma = 1$); and apply the Sampling theorem to write $f(x)$ in terms of an infinite series of sinc functions.

Exercise 6 Recall the function $\widehat{h}_\eta(\omega)$ in (7.2.13) and its inverse Fourier transform (IFT) $h_\eta(x)$ in (7.2.16).

(a) For a constant $c > 0$, what is the bandwidth of the function $k_c(x) = h_\eta(cx)$?
(b) Evaluate $k_c(0)$ by applying L'Hospital's rule.
(c) For $0 < c \leq 1$, verify the Sampling theorem for $f(t) = k_c(t)$, by allowing $\eta = 1$.
(d) For $c = \frac{1}{m}$, where m is a positive integer, verify the Sampling theorem directly.

Exercise 7 Apply (7.2.17) to verify (7.2.20).

7.3 Isotropic Diffusion PDE

The study of (heat) diffusion on a bounded region D in $\mathbb{R}^s$, for any dimension $s \geq 1$, has already been studied in Sect. 6.5 of Chap. 6. When D is perfectly insulated on the boundary ∂D of D, the mathematical model of the diffusion problem is the partial differential equation (PDE)

$$\frac{\partial}{\partial t} u(\mathbf{x}, t) = c\nabla^2 u(\mathbf{x}, t), \ \mathbf{x} \in D, t \geq 0,$$

with zero Neumann's boundary condition, namely:

$$\frac{\partial}{\partial \mathbf{n}} u(\mathbf{x}, 0) = 0, \ \mathbf{x} \in \partial D,$$

where $\mathbf{n} = \mathbf{n_x}$ denotes the unit outer normal vector at $\mathbf{x} \in \partial D$. Here and throughout, recall that the $c > 0$ is a positive constant, called heat conductivity. In this section, we consider the same PDE, but on the entire space $\mathbb{R}^s$, $s \geq 1$, instead of a bounded region D. Hence, there is no boundary condition; and with initial heat content $u(\mathbf{x}, 0) = u_0(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^s$, at $t = 0$, the mathematical model is the initial-value PDE

$$\begin{cases} \dfrac{\partial}{\partial t} u(\mathbf{x}, t) = c\nabla^2 u(\mathbf{x}, t), & \mathbf{x} \in \mathbb{R}^s, t \geq 0, \\[2mm] u(\mathbf{x}, 0) = u_0(\mathbf{x}), & \mathbf{x} \in \mathbb{R}^s. \end{cases} \tag{7.3.1}$$

Our mathematical tool is convolution with the Gaussian function

$$g_\sigma(\mathbf{x}) = g_\sigma(x_1) \cdots g_\sigma(x_s),$$

where $\mathbf{x} = (x_1, \ldots, x_s) \in \mathbb{R}^s$. Let us first focus on heat diffusion on $\mathbb{R} = \mathbb{R}^1$, with given initial heat content $u_0(x)$, $-\infty < x < \infty$.

Recall from Sect. 7.1 that for any constant $\sigma > 0$, the Gaussian function $g_\sigma(x)$ is defined by

$$g_\sigma(x) = \frac{1}{2\sigma\sqrt{\pi}} e^{-\left(\frac{x}{2\sigma}\right)^2}$$

as in (7.1.13) on p.324, satisfies

$$\int_{-\infty}^{\infty} g_\sigma(x)dx = 1, \ \text{all} \ \sigma > 0$$

(see (7.1.14) on p.324), and its Fourier transform is given by

$$\widehat{g}_\sigma(\omega) = e^{-\sigma^2\omega^2}, \qquad (7.3.2)$$

as shown in (7.1.20) of Theorem 3 in Sect. 7.1, on p. 326. Let $c > 0$ be the conductivity constant in the PDE (7.3.1) and introduce the "time" parameter

$$t = \frac{\sigma^2}{c} \text{ or } \sigma^2 = ct \qquad (7.3.3)$$

to define the function $G(x, t)$ of two variables:

$$G(x, t) = g_\sigma(x) = g_{\sqrt{ct}}(x) = \frac{t^{-\frac{1}{2}}}{2\sqrt{\pi c}} e^{-\frac{x^2}{4c}t^{-1}}, \qquad (7.3.4)$$

with $x \in \mathbb{R}$ to be called the spatial variable, and $t \geq 0$ to be called the time variable. Then $G(x, t)$ satisfies the PDE (7.3.1) in that

$$\frac{\partial}{\partial t}G(x, t) = c\frac{\partial^2}{\partial x^2}G(x, t), \ x \in \mathbb{R}, t > 0. \qquad (7.3.5)$$

Indeed, by taking the partial derivatives of $G(x, t)$ in (7.3.4), we have

$$\frac{\partial}{\partial t}G(x, t) = \frac{1}{2\sqrt{\pi c}} e^{-\frac{x^2}{4c}t^{-1}}\left\{-\frac{1}{2}t^{-\frac{3}{2}} + t^{-\frac{1}{2}}(\frac{x^2}{4c})t^{-2}\right\};$$

$$\frac{\partial}{\partial x}G(x, t) = \frac{t^{-\frac{1}{2}}}{2\sqrt{\pi c}} e^{-\frac{x^2}{4c}t^{-1}}\left\{-\frac{t^{-1}}{2c}x\right\},$$

so that

$$\begin{aligned}
\frac{\partial^2}{\partial x^2}G(x, t) &= \frac{t^{-\frac{1}{2}}}{2\sqrt{\pi c}} e^{-\frac{x^2}{4c}t^{-1}}\left\{-\frac{t^{-1}}{2c} + (-\frac{t^{-1}}{2c}x)^2\right\} \\
&= \frac{1}{c}\frac{1}{2\sqrt{\pi c}} e^{-\frac{x^2}{4c}t^{-1}}\left\{-\frac{1}{2}t^{-\frac{3}{2}} + (\frac{x^2}{4c})t^{-\frac{1}{2}} \cdot t^{-2}\right\} \\
&= \frac{1}{c}\frac{\partial}{\partial t}G(x, t),
\end{aligned}$$

which proves (7.3.5). Next, it follows from Theorem 5 on p.327 in Sect. 7.1 that

$$G(x, 0) = \delta(x), \ x \in \mathbb{R},$$

where $\delta(x)$ denotes the "Dirac delta" distribution (also commonly called the "delta function") in that if $f \in L_2(\mathbb{R})$ is continuous at x, then $f(x)$ is "reproduced" by convolution with the delta function:

$$f * \delta(x) = f(x),$$

meaning that

$$\lim_{t \to 0^+} \int_{-\infty}^{\infty} f(y)G(x - y, t)dy = f(x), \ x \in \mathbb{R} \tag{7.3.6}$$

(see Theorem 5 on p.327 in Sect. 7.1).

Remark 1 We have assumed that $f \in L_\infty(\mathbb{R})$ in Theorem 5 on p.327 for (7.3.6) to hold, but it was only for simplicity of the proof. In fact, since for any fixed $t > 0$, $f(y)G(x - y, t)$ is integrable for all $f \in PC(\mathbb{R})$ with "at most polynomial growth", meaning that

$$f(x)x^{-n} \in L_\infty(\mathbb{R})$$

for some integer $n > 0$, the statement of Theorem 5 on p.327 (or equivalently (7.3.6)) remains valid for all $f \in PC(\mathbb{R})$ with at most polynomial growth. ∎

In other words, we have the following result.

Theorem 1 $\boxed{\textbf{Solution of diffusion PDE for } \mathbb{R}^1}$ *Let $u_0 \in PC(\mathbb{R})$ with at most polynomial growth. Then the solution of the initial value PDE*

$$\begin{cases} \dfrac{\partial}{\partial t}u(x, t) = c\dfrac{\partial^2}{\partial x^2}u(x, t), & x \in \mathbb{R}, t \geq 0, \\[3mm] u(x, 0) = u_0(x), & x \in \mathbb{R}, \end{cases} \tag{7.3.7}$$

is given by

$$u(x, t) = \int_{-\infty}^{\infty} u_0(y)G(x - y, t)dy. \tag{7.3.8}$$

The proof of (7.3.8) has already been discussed in (7.3.6) with $f(x) = u_0(x)$ and Remark 1. The idea is that if the heat source (i.e. initial heat content) is the delta function $\delta(x)$, then the heat distribution for $t > 0$ is the Gaussian function $G(x, t)$, as shown at the top of Fig. 7.1.

To show that the heat distribution $u(x, t)$, with initial heat content $u_0(x)$, is the solution of the initial PDE (7.3.7), we simply apply (7.3.5) to obtain

$$\begin{aligned} \frac{\partial}{\partial t}u(x, t) &= \frac{\partial}{\partial t}\int_{-\infty}^{\infty} u_0(y)G(x - y, t)dy \\[2mm] &= \int_{-\infty}^{\infty} u_0(y)\frac{\partial}{\partial t}G(x - y, t)dy \\[2mm] &= \int_{-\infty}^{\infty} u_0(y)\left\{c\frac{\partial^2}{\partial x^2}G(x - y, t)\right\}dy \end{aligned}$$
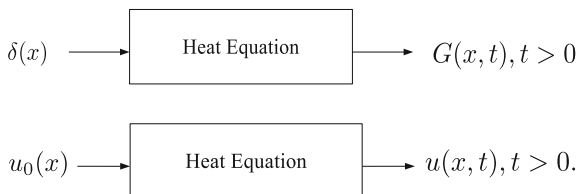
Fig. 7.1. Diffusion with delta heat source (*top*) and arbitrary heat source (*bottom*)

$$= c \frac{\partial^2}{\partial x^2} \int_{-\infty}^{\infty} u_0(y) G(x - y, t) dy$$

$$= c \frac{\partial^2}{\partial x^2} u(x, t),$$

by the definition of $u(x, t)$ in (7.3.8). Hence, $u(x, t)$ as defined by (7.3.8) is the output as shown in the bottom of Fig. 7.1, with input $u_0(x)$. ∎

Example 1 Compute the solution $u(x, t)$ of the initial value (heat diffusion) PDE in (7.3.7), with initial (or input) function

$$u_0(x) = a_\alpha \cos \alpha x + b_\alpha \sin \alpha x,$$

where α is any real number and a_α, b_α are arbitrary constants.

Solution By Theorem 1, the solution is obtained by computing the convolution of $u_0(x)$ with the Gaussian function $G(x, t)$ with respect to the spatial variable x, where $t \geq 0$ is fixed while the convolution operation is performed. Although there is no need to consider the following three separate cases, we will do so in this first example to show the computational steps more transparently.

Case 1. For $\alpha = 0$, the input function $u_0(x) = a_0$ is a constant. Hence, $u_0(x - y) = a_0$ and

$$u(x, t) = \left(u_0 * G(\cdot, t) \right)(x) = \int_{-\infty}^{\infty} a_0 g_\sigma(y) dy = a_0,$$

since the Gaussian $g_\sigma(x)$ is normalized with integral over $(-\infty, \infty)$ equal to 1.

Case 2. For $b_\alpha = 0$, the initial function is

$$u_0(x) = a_\alpha \cos \alpha x = \frac{a_\alpha}{2} \left(e^{i\alpha x} + e^{-i\alpha x} \right).$$

Hence, for fixed $\sigma^2 = ct$, the convolution becomes

$$u(x, t) = \left(u_0 * g_\sigma \right)(x)$$

$$
\begin{aligned}
&= \frac{a_\alpha}{2}\left(\int_{-\infty}^{\infty} e^{i\alpha(x-y)}\, g_\sigma(y)dy + \int_{-\infty}^{\infty} e^{-i\alpha(x-y)}\, g_\sigma(y)dy \right) \\
&= \frac{a_\alpha}{2}\left(e^{i\alpha x}\widehat{g_\sigma}(\alpha) + e^{-i\alpha x}\widehat{g_\sigma}(-\alpha) \right) \\
&= \frac{a_\alpha}{2}\left(e^{i\alpha x}e^{-\sigma^2\alpha^2} + e^{-i\alpha x}e^{-\sigma^2(-\alpha)^2} \right) \\
&= \frac{a_\alpha}{2}\left(e^{i\alpha x} + e^{-i\alpha x} \right)e^{-\sigma^2\alpha^2}
\end{aligned}
$$

by (7.3.2). Since $\sigma^2 = ct$, we have

$$
u(x,t) = a_\alpha e^{-c\alpha^2 t}\, \cos \alpha x.
$$

Case 3. For $a_\alpha = 0$, the initial function is

$$
u_0(x) = b_\alpha \sin \alpha x = \frac{b_\alpha}{2i}\left(e^{i\alpha x} - e^{-i\alpha x} \right).
$$

Therefore, the same computation as above yields

$$
\begin{aligned}
u(x,t) &= \frac{b_\alpha}{2i}\left(e^{i\alpha x} - e^{-i\alpha x} \right)e^{-\sigma^2\alpha^2} \\
&= b_\alpha e^{-c\alpha^2 t}\, \sin \alpha x,
\end{aligned}
$$

since $\sigma^2 = ct$.

Combining the above computational results, we obtain the solution of the initial value PDE (7.3.7):

$$
u(x,t) = e^{-c\alpha^2 t}\left(a_\alpha \cos \alpha x + b_\alpha \sin \alpha x \right) = e^{-c\alpha^2 t}u_0(x)
$$

for all $x \in \mathbb{R}$ and $t \geq 0$. ∎

In general, if the initial (input) function $u_0(x)$ is in $PC_{2\pi}^\star$, the same computational steps apply to yield the solution $u(x,t)$ of the initial value PDE (7.3.7), as follows.

Example 2 *Let $u_0 \in PC[0, M]$, $M > 0$ such that the partial sums $(S_n u_0)(x)$ of the Fourier series of $u_0(x)$ are uniformly bounded. Show that the solution $u(x,t)$ of the initial value PDE (7.3.7) with initial heat content $u_0(x)$ is given by*

$$
u(x,t) = \frac{a_0}{2} + \sum_{k=1}^{\infty} e^{-c\left(\frac{k\pi}{M}\right)^2 t}\left(a_k \cos \frac{k\pi}{M}x + b_k \sin \frac{k\pi}{M}x \right), \tag{7.3.9}
$$

where

$$a_k = \frac{1}{M} \int_{-M}^{M} u_0(x) \cos \frac{k\pi}{M} x \, dx, \ k = 0, 1, 2, \ldots$$

$$b_k = \frac{1}{M} \int_{-M}^{M} u_0(x) \sin \frac{k\pi}{M} x \, dx, \ k = 1, 2, \ldots.$$

Solution Consider the Fourier cosine and sine series expansion of $u_0(x)$ in Theorem 2 on p.278 of Chap. 6, with $d = M$. Since $(S_n u_0)(x)$ are uniformly bounded, we may apply Lebesgue's dominated convergence theorem in Theorem 4 of the Appendix (or Sect. 7.6) of this chapter to interchange summation and integration, namely:

$$
\begin{aligned}
u(x, t) &= \big(u_0 * G(\cdot, t)\big)(x) \\
&= \frac{a_0}{2} + \sum_{k=1}^{\infty} \left(\frac{a_k - ib_k}{2} \int_{-\infty}^{\infty} e^{i\frac{k\pi}{M}(x-y)} G(y, t) dy \right. \\
&\qquad\qquad \left. + \frac{a_k + ib_k}{2} \int_{-\infty}^{\infty} e^{-i\frac{k\pi}{M}(x-y)} G(y, t) dt \right) \\
&= \frac{a_0}{2} + \sum_{k=1}^{\infty} \left(\frac{a_k - ib_k}{2} e^{i\frac{k\pi}{M}x} \widehat{g}_\sigma(k) + \frac{a_k + ib_k}{2} e^{-i\frac{k\pi}{M}x} \widehat{g}_\sigma(-k) \right) \\
&= \frac{a_0}{2} + \sum_{k=1}^{\infty} \left(a_k \frac{e^{i\frac{k\pi}{M}x} + e^{-i\frac{k\pi}{M}x}}{2} + b_k \frac{e^{i\frac{k\pi}{M}x} - e^{-i\frac{k\pi}{M}x}}{-2i} \right) \widehat{g}_\sigma(k) \\
&= \frac{a_0}{2} + \sum_{k=1}^{\infty} e^{-\sigma^2 (\frac{k\pi}{M})^2} \left(a_k \cos \frac{k\pi}{M} x + b_k \sin \frac{k\pi}{M} x \right) \\
&= \frac{a_0}{2} + \sum_{k=1}^{\infty} e^{-(\frac{k\pi}{M})^2 ct} \left(a_k \cos \frac{k\pi}{M} t + b_k \sin \frac{k\pi}{M} t \right),
\end{aligned}
$$

since $\sigma^2 = ct$, where again the formula $\widehat{g}_\sigma(\omega) = e^{-\sigma^2 \omega^2}$ in (7.3.2) is applied. ∎

Example 3 Find the solution $u(x, t)$ of the initial value (heat diffusion) PDE (7.3.7) with initial heat content $u_0(x) = x^n$, for $n = 1$ and 2.

Solution Since the PDE (7.3.7) describes the heat diffusion process, let us consider $u_0(x)$ as the initial temperature at $x \in \mathbb{R}$. Hence, the solution $u(x, t)$ is the temperature at the time instant $t > 0$, at the same position $x \in \mathbb{R}$.

For $n = 1$, the temperature at $t > 0$ and location $x \in \mathbb{R}$ is given by

$$
\begin{aligned}
u(x, t) &= \int_{-\infty}^{\infty} (x - y) g_\sigma(y) dy \\
&= x \int_{-\infty}^{\infty} g_\sigma(y) dy - \int_{-\infty}^{\infty} y g_\sigma(y) dy,
\end{aligned}
$$

where $\sigma^2 = ct$. Since the first integral is equal to 1 and the second integral is 0 (with odd function $y g_\sigma(y)$), we have

$$u(x, t) = x, \text{ for all } t \geq 0.$$

That is, the temperature does not change with time; or there is no diffusion at all.

For $n = 2$, since $(x - y)^2 = x^2 - 2xy + y^2$, the same argument as above yields

$$u(x, t) = \int_{-\infty}^{\infty} (x - y)^2 g_\sigma(y) dy = x^2 + \int_{-\infty}^{\infty} y^2 g_\sigma(y) dy,$$

where $\sigma^2 = ct$. Observe that because

$$\int_{-\infty}^{\infty} y^2 e^{-\alpha y^2} dx = -\frac{\partial}{\partial \alpha} \int_{-\infty}^{\infty} e^{-\alpha y^2} dy$$

$$= -\frac{\partial}{\partial \alpha} \sqrt{\frac{\pi}{\alpha}} = \frac{1}{2} \sqrt{\pi} \alpha^{-3/2},$$

we obtain

$$\int_{-\infty}^{\infty} y^2 g_\sigma(y) dy = \frac{1}{2\sigma\sqrt{\pi}} \int_{-\infty}^{\infty} y^2 e^{-\left(\frac{y}{2\sigma}\right)^2} dy$$

$$= \frac{1}{2\sigma\sqrt{\pi}} \frac{1}{2} \sqrt{\pi} (2\sigma)^3 = 2\sigma^2 = 2ct,$$

and this yields:

$$u(x, t) = x^2 + 2ct, \text{ for all } t > 0.$$

Observe that the temperature at x increases from $u(x, 0) = x^2$ to $u(x, c) = x^2 + 2ct$ as $t > 0$ increases. This is truly "global warming" everywhere! ∎

Let us now extend our discussion from one spatial variable $x \in \mathbb{R}$ to s spatial variables $\mathbf{x} = (x_1, \ldots, x_s) \in \mathbb{R}^s$. Let $|\mathbf{x}|$ denote the Euclidean norm of $\mathbf{x} \in \mathbb{R}^s$; that is

$$|\mathbf{x}|^2 = x_1^2 + \cdots + x_s^2,$$

and observe that the Gaussian function $g_\sigma(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^s$, defined by

$$g_\sigma(\mathbf{x}) = \frac{1}{(4\pi\sigma^2)^{s/2}} e^{-\frac{|\mathbf{x}|^2}{4\sigma^2}}, \tag{7.3.10}$$

can be written as the product of the 1-dimensional Gaussian functions; namely

$$g_\sigma(\mathbf{x}) = g_\sigma(x_1, \ldots, x_s) = g_\sigma(x_1) \cdots g_\sigma(x_s).$$

Hence, when the time variable t is defined by (7.3.3); that is $\sigma^2 = ct$, the extension of $G(x, t)$ in (7.3.4) to $\mathbb{R}^s$, $s \geq 2$, is given by

$$G(\mathbf{x}, t) = G(x_1, \ldots, x_s, t) = g_\sigma(\mathbf{x}) = \frac{t^{-\frac{s}{2}}}{(4\pi c)^{s/2}} e^{-\frac{|\mathbf{x}|^2}{4c} t^{-1}}. \qquad (7.3.11)$$

Indeed, by (7.3.10) and (7.3.4), we have

$$\begin{aligned}
G(\mathbf{x}, t) &= g_\sigma(x_1) \cdots g_\sigma(x_s) \\
&= \left(\frac{1}{\sqrt{4\pi\sigma^2}} e^{-x_1^2/4\sigma^2} \right) \cdots \left(\frac{1}{\sqrt{4\pi\sigma^2}} e^{-x_s^2/4\sigma^2} \right) \\
&= \frac{1}{(4\pi\sigma^2)^{s/2}} e^{-x_1^2/4\sigma^2} \cdots e^{-x_s^2/4\sigma^2} \\
&= \frac{1}{(4\pi\sigma^2)^{s/2}} e^{-(x_1^2 + \cdots + x_s^2)/4\sigma^2} = \frac{1}{(4\pi\sigma^2)^{s/2}} e^{-|\mathbf{x}|^2/4\sigma^2} \\
&= \frac{t^{-\frac{s}{2}}}{(4\pi c)^{s/2}} e^{-\frac{|\mathbf{x}|^2}{4c} t^{-1}},
\end{aligned}$$

as desired.

Now, let us consider the s-dimensional convolution defined by

$$(u_0 * h)(\mathbf{x}) = \int_{\mathbb{R}^s} u_0(\mathbf{y}) h(\mathbf{x} - \mathbf{y}) d\mathbf{y}. \qquad (7.3.12)$$

Then for fixed $t \geq 0$, the s-dimensional convolution with the spatial variables $\mathbf{x} = (x_1, \ldots, x_s)$ of $h(\mathbf{x}) = G_c(\mathbf{x}, t)$ can be written as consecutive 1-dimensional convolutions; namely,

$$\begin{aligned}
\left(u_0 * G(\cdot, t) \right)(\mathbf{x}) &= \int_{\mathbb{R}^s} u_0(\mathbf{y}) G(\mathbf{x} - \mathbf{y}, t) d\mathbf{y} \\
&= \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{s \text{ integrals}} u_0(y_1, \ldots, y_s) G(x_1 - y_1, t) \cdots G(x_s - y_s, t) dy_1 \cdots dy_s \\
&= \left(u_0 \underbrace{(., ., \ldots, .)}_{s \text{ components}} *_1 g_\sigma *_2 \cdots *_s g_\sigma \right)(x_1, \ldots, x_s), \qquad (7.3.13)
\end{aligned}$$

where "$*_k$" denotes the 1-dimensional convolution with respect to the kth component (say y_k in (7.3.12)).

Hence, as in the 1-dimensional case, it is not difficult to verify that

$$G(\mathbf{x}, t) = G(x_1, \ldots, x_s, t)$$

is the solution of the s-dimensional heat equation with $\delta(\mathbf{x}) = \delta(x_1) \cdots \delta(x_s)$ as the initial heat source; namely,

$$
\begin{cases}
\dfrac{\partial}{\partial t} G(\mathbf{x}, t) = c \nabla^2 G(\mathbf{x}, t), & \mathbf{x} \in \mathbb{R}^s, t \geq 0, \\[2ex]
G(\mathbf{x}, 0) = \delta(\mathbf{x}) = \delta(x_1) \cdots \delta(x_s), & \mathbf{x} \in \mathbb{R}^s,
\end{cases}
\tag{7.3.14}
$$

where ∇^2 denotes the Laplace operator, defined by

$$
\nabla^2 G(\mathbf{x}, t) = \frac{\partial^2}{\partial x_1^2} G(\mathbf{x}, t) + \cdots + \frac{\partial^2}{\partial x_s^2} G(\mathbf{x}, t).
$$

To verify that $G(\mathbf{x}, t)$ satisfies the heat diffusion equation, we simply follow the same computations in the derivation of (7.3.5), as follows:

$$
\frac{\partial}{\partial t} G(\mathbf{x}, t) = \frac{1}{(4\pi c)^{s/2}} e^{-(\frac{|\mathbf{x}|^2}{4c})t^{-1}} \left\{ -\frac{s}{2} t^{-\frac{s+2}{2}} + t^{-\frac{s}{2}} (\frac{|\mathbf{x}|^2}{4c}) t^{-2} \right\};
$$

$$
\frac{\partial}{\partial x_k} G(x_1, \ldots, x_s, t) = \frac{t^{-\frac{s}{2}}}{(4\pi c)^{s/2}} e^{-\frac{|\mathbf{x}|^2}{4c} t^{-1}} \left\{ -\frac{t^{-1}}{2c} x_k \right\},
$$

so that

$$
\frac{\partial^2}{\partial x_k^2} G(\mathbf{x}, t) = \frac{t^{-\frac{s}{2}}}{(4\pi c)^{s/2}} e^{-\frac{|\mathbf{x}|^2}{4c} t^{-1}} \left\{ -\frac{t^{-1}}{2c} + (-\frac{t^{-1}}{2c} x_k)^2 \right\}.
$$

Hence, it follows that

$$
\begin{aligned}
c \nabla^2 G(\mathbf{x}, t) &= c \sum_{k=1}^{s} \frac{\partial^2}{\partial x_k^2} G(x_1, \ldots, x_s, t) \\
&= \frac{t^{-\frac{s}{2}}}{(4\pi c)^{s/2}} e^{-\frac{|\mathbf{x}|^2}{4c} t^{-1}} \left\{ -\frac{s t^{-1}}{2c} + \frac{t^{-2}}{4c^2} \sum_{k=1}^{s} x_k^2 \right\} \\
&= \frac{1}{(4\pi c)^{s/2}} e^{-\frac{|\mathbf{x}|^2}{4c} t^{-1}} \left\{ -\frac{s}{2} t^{-\frac{s}{2}-1} + t^{-\frac{s}{2}} (\frac{|\mathbf{x}|^2}{4c})^2 t^{-2} \right\} \\
&= \frac{\partial}{\partial t} G(\mathbf{x}, t).
\end{aligned}
$$

To apply the above result to an arbitrary heat source, we simply follow the same argument for the 1-dimensional setting and apply (7.3.13) and (7.3.14) to obtain the solution of the initial value PDE:

$$\begin{cases} \dfrac{\partial}{\partial t} u(\mathbf{x}, t) = c\nabla^2 u(\mathbf{x}, t), & \mathbf{x} \in \mathbb{R}^s, t \geq 0, \\[3mm] u(\mathbf{x}, 0) = u_0(\mathbf{x}), & \mathbf{x} \in \mathbb{R}^s, \end{cases} \tag{7.3.15}$$

where the initial condition is any "reasonable measurable function" $u_0(\mathbf{x}) = u_0(x_1, \ldots, x_s)$ defined on $\mathbb{R}^s$.

We summarize the above discussion in the following theorem.

Theorem 2 $\boxed{\textbf{Solution of diffusion PDE for } \mathbb{R}^s}$ *Let $u_0(\mathbf{x})$ be a measurable function in $\mathbb{R}^s$, $s \geq 1$, with at most polynomial growth such that the set of $\mathbf{x} \in \mathbb{R}^s$ on which $u_0(\mathbf{x})$ is discontinuous has (Lebesgue) measure zero. Then the solution of the initial value PDE (7.3.15) is given by*

$$u(\mathbf{x}, t) = \big(u_0 * G(\cdot, t)\big)(\mathbf{x}), \tag{7.3.16}$$

as defined in (7.3.13).

Example 4 Let $c > 0$ and ∇^2 be the Laplace operator defined by

$$\nabla^2 f(x, y) = \frac{\partial^2}{\partial x^2} f(x, y) + \frac{\partial^2}{\partial y^2} f(x, y).$$

Re-formulate the solution $u(x, y, t)$ of the 2-dimensional (heat diffusion) initial value PDE

$$\frac{\partial}{\partial t} u(x, y, t) = c\nabla^2 u(x, y, t)$$

in (7.3.15) with initial (input) function $u_0(x, y)$, according to Theorem 2 explicitly in terms of the 1-dimensional Gaussian function.

Solution According to Theorem 2 and the commutative property of the convolution operation, we have

$$\begin{aligned} u(x, y, t) \\ = \int_{-\infty}^{\infty} g_\sigma(y_1) \Big\{ \int_{-\infty}^{\infty} u_0(x - x_1, y - y_1) g_\sigma(x_1) dx_1 \Big\} dy_1 \qquad (7.3.17) \\ = \frac{t^{-1}}{4\pi c} \int_{-\infty}^{\infty} e^{-\frac{t^{-1}}{4c} y_1^2} \Big\{ \int_{-\infty}^{\infty} u_0(x - x_1, y - y_1) e^{-\frac{t^{-1}}{4c} x_1^2} dx_1 \Big\} dy_1, \end{aligned}$$

where $\sigma^2 = ct$. ∎

Definition 1 *A function $u_0(x, y)$ of two variables is said to be separable, if there exist functions f_j and h_k of one variable, such that*

$$u_0(x, y) = \sum_{j,k} a_{j,k} f_j(x) h_k(y) \tag{7.3.18}$$

for some constants $a_{j,k}$, where $\sum_{j,k}$ denotes a finite double sum, such as:

$$\sum_{j=0}^{m} \sum_{k=0}^{n}, \quad \sum_{\ell=0}^{n} \sum_{j+k=\ell}, \quad \sum_{j=0}^{m} \sum_{k=0}^{n-j}, \quad etc.$$

Example 5 Let $u_0(x, y)$ be a separable function as defined by (7.3.18). Write out the solution $u(x, y, t)$ of the 2-dimensional (heat diffusion) PDE in Example 4 with initial condition $u_0(x, y)$ in the most useful form for computation.

Solution Let $\sigma^2 = ct$ and apply the formulation (7.3.17) to obtain

$$
\begin{aligned}
u(x, y, t) \\
&= \int_{-\infty}^{\infty} g_\sigma(y_1) \left\{ \int_{-\infty}^{\infty} \sum_{j,k} a_{j,k} f_j(x - x_1) h_k(y - y_1) g_\sigma(x_1) dx_1 \right\} dy_1 \\
&= \sum_{j,k} a_{j,k} \left(\int_{-\infty}^{\infty} f_j(x - x_1) g_\sigma(x_1) dx_1 \right) \tag{7.3.19} \\
&\quad \times \left(\int_{-\infty}^{\infty} h_k(y - y_1) g_\sigma(y_1) dy_1 \right),
\end{aligned}
$$

with $\sigma = \sqrt{ct}$. After computing (7.3.19), we may write out the solution $u(x, y, t)$ by replacing σ with $\sqrt{c}\, t^{1/2}$, as follows:

$$
\begin{aligned}
u(x, y, t) &= \frac{t^{-1}}{4\pi c} \sum_{j,k} a_{j,k} \left(\int_{-\infty}^{\infty} f_j(x - x_1) e^{-\frac{t^{-1}}{4c} x_1^2} dx_1 \right) \\
&\quad \times \left(\int_{-\infty}^{\infty} h_k(y - y_1) e^{-\frac{t^{-1}}{4c} y_1^2} dy_1 \right).
\end{aligned}
$$

∎

Example 6 Apply the solution in Example 5 to compute the solution $u(x, y, t)$ of the 2-dimensional (heat diffusion) initial value PDE in Example 4 with initial condition $u_0(x, y)$ given by (7.3.18), where $f_j(x) = \cos \frac{j\pi}{M} x$ and $h_k(y) = \cos \frac{k\pi}{N} y$ for $M, N > 0$.

Solution For $f_j(x) = \cos \frac{j\pi}{M} x$ and $h_k(y) = \cos \frac{k\pi}{N} y$, application of the solution in Example 1 yields

$$\int_{-\infty}^{\infty} f_j(x - x_1) g_\sigma(x_1) dx_1 = e^{-\sigma^2 (\frac{j\pi}{M})^2} \cos \frac{j\pi}{M} x = e^{-(\frac{j\pi}{M})^2 ct} \cos \frac{j\pi}{M} x.$$

Hence, it follows from (7.3.19) in the previous example that

$$u(x, y, t) = \sum_{j,k} a_{j,k} e^{-\left((\frac{j\pi}{M})^2+(\frac{k\pi}{N})^2\right)ct} \cos \frac{j\pi}{M} x \cos \frac{k\pi}{M} y.$$

∎

Exercises

Exercise 1 Solve the initial value PDE

$$\frac{\partial}{\partial t} u(x, t) = c \frac{\partial^2}{\partial x^2} u(x, t), \; x \in \mathbb{R}, t \geq 0,$$

with initial value $u(x, 0) = u_0(x)$, where $c > 0$ is a constant and

$$u_0(x) = 10 - \cos 2x + 5 \sin 3x.$$

Exercise 2 Repeat Exercise 1 by considering the initial function $u_0(x) = \sin^3 x$.

Exercise 3 Repeat Exercise 1 by considering the initial function $u_0(x) = x^n$, where $n \geq 3$ is any positive integer.

Exercise 4 Solve the 2-dimensional initial value PDE

$$\frac{\partial}{\partial t} u(x, y, t) = c \nabla^2 u(x, y, t), \; (x, y) \in \mathbb{R}^2, t \geq 0,$$

where $c > 0$ is a constant, with initial condition

$$u(x, y, 0) = u_0(x, y) = \sum_{j,k} a_{j,k}(\cos jx)(\sin ky),$$

for some constants $a_{j,k}$.

Exercise 5 Repeat Exercise 4 by considering the initial condition

$$u_0(x, y) = \sum_{j,k} b_{j,k}(\sin jx)(\sin ky)$$

for some constants $b_{j,k}$.

Exercise 6 Repeat Exercise 4 by considering the initial condition

$$u_0(x, y) = a_{0,0} + a_{1,0}x + a_{0,1}y + a_{2,0}x^2 + a_{1,1}xy + a_{0,2}y^2 + a_{3,0}x^3$$
$$+ a_{2,1}x^2 y + a_{1,2}xy^2 + a_{0,3}y^3,$$

for some constants $a_{j,k}$, $j + k = \ell$, where $\ell = 0, 1, 2, 3$.

7.4 Time-Frequency Localization

In Sect. 5.3 of Chap. 5, we have discussed the application of DCT to image sub-blocks, as opposed to the entire digital image. This is essential since multiplication of high-dimensional matrices is not only computational expensive, but also requires significantly more memory. For example, for JPEG compression, DCT is applied to 8×8 pixel blocks.

To extend this idea to the computation of the Fourier transform, if the function $f(x)$ is truncated by some characteristic function $\chi_{(a,b)}(x)$, then computation of

$$\left(f \chi_{(a,b)}\right)^{\wedge}(\omega) = \int_a^b f(x)e^{-i\omega x}dx$$

is certainly much simpler than that of $\widehat{f}(\omega)$. In general, a more desirable window function $u(x)$ could be used in place of the characteristic function $\chi_{(a,b)}(x)$, and this window should be allowed to move (continuously) along the x-axis, instead of partitioning the x-axis into disjoint intervals. This is the central idea of the so-called "short-time" Fourier transform. Since this transform localizes the function $f(x)$ before the Fourier transform is applied, we will also call it localized Fourier transform, as follows.

Definition 1 | **Short-time Fourier transform** | *Let $u \in \left(L_1 \cap L_2\right)(\mathbb{R})$ and $x \in \mathbb{R}$. Then for any $f \in L_2(\mathbb{R})$, the integral transform*

$$\left(\mathbb{F}_u f\right)(x, \omega) = \int_{-\infty}^{\infty} f(t)u(t-x)e^{-i\omega t}dt \qquad (7.4.1)$$

is called the localized Fourier transform (LFT), or short-time Fourier transform (STFT) of the function $f(x)$ at the time-frequency (or space-frequency) point $(x, \omega) \in \mathbb{R}^2$.

Remark 1   In contrast with the Fourier transform $\mathbb{F}$ that takes a function $f(x)$ from the time (or spatial) domain $\mathbb{R}$ to $\widehat{f}(\omega)$ in the frequency domain $\mathbb{R}$, the LFT $\mathbb{F}_u$, with "window function" $u(x)$, takes $f(x)$ from the time (or spatial) domain $\mathbb{R}$ to the time-frequency domain $\mathbb{R}^2$. For this reason, we use t (instead of x) as the dummy variable for the integration in (7.4.1), while reserving the variable x in the time-frequency coordinate $(x, \omega) \in \mathbb{R}^2$. ∎

To localize the inverse Fourier transform (see Theorem 4. on p.332) with some window function $v \in \left(L_1 \cap L_2\right)(\mathbb{R})$, we adopt the notation $\mathbb{F}^{\#}$ in (7.2.9) on p.332 to introduce the localized inverse Fourier transform $\mathbb{F}_v^{\#}$ as follows.

Definition 2 | **Localized inverse Fourier transform** | *Let $v \in \left(L_1 \cap L_2\right)(\mathbb{R})$ and $\omega \in \mathbb{R}$. Then for any $f(x) \in L_2(\mathbb{R})$ with Fourier transform $\widehat{f}(\omega)$, the integral transform*

$$\left(\mathbb{F}_v^{\#}\widehat{f}\right)(x, \omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \widehat{f}(\xi) v(\xi - \omega) e^{ix\xi} d\xi \tag{7.4.2}$$

is called the localized inverse Fourier transform (LIFT) of $\widehat{f}(\omega)$ at the time-frequency (or spatial-frequency) point $(x, \omega) \in \mathbb{R}^2$.

Remark 2   As mentioned in Remark 1, since ω is reserved for the frequency variable of the time-frequency coordinate $(x, \omega) \in \mathbb{R}^2$, the dummy variable ξ is used for the integration in (7.4.2). We also remark that in view of Theorem 1 below and the recovery formula for $f(x)$ in (7.4.12) of Theorem 4 on p.332 to be discussed in this section, the notion of LIFT is justified.   ∎

To quantify the localization properties of the LFT and LIFT, we introduce the notion of "window width" in the following.

Definition 3 $\boxed{\textbf{Time-frequency window width}}$ *Let $u \in (L_1 \cap L_2)(\mathbb{R})$ be a nonzero function on $\mathbb{R}$ such that $xu(x) \in L_2(\mathbb{R})$. Then*

$$x^{\star} = \frac{1}{\|u\|_2^2} \int_{-\infty}^{\infty} x|u(x)|^2 dx \tag{7.4.3}$$

is called the center of the localization window function $u(x)$, and

$$\Delta_u = \left\{ \frac{1}{\|u\|_2^2} \int_{-\infty}^{\infty} (x - x^{\star})^2 |u(x)|^2 dx \right\}^{1/2} \tag{7.4.4}$$

is called the radius of $u(x)$. In addition, the window width of $u(x)$ is defined by $2\Delta_u$.

Observe that for $u \in L_2(\mathbb{R})$, if $xu(x) \in L_2(\mathbb{R})$, then $xu(x)^2 \in L_1(\mathbb{R})$ (see Exercise 2). Thus the center $x^{\star}$ is well-defined.

In the following, we will see that if $u \in (L_1 \cap L_2)(\mathbb{R})$ with Fourier transform $\widehat{u}(\omega)$ also in $(L_1 \cap L_2)(\mathbb{R})$, then by choosing $v(\omega) = (\overline{\mathbb{F}u})(\omega) = \widehat{u}(-\omega)$ (see Exercise 3) for the LIFT $\mathbb{F}_v^{\#}$ in (7.4.2), we have simultaneous time and frequency localization.

Theorem 1 $\boxed{\textbf{Time-frequency localization}}$ *Let $u \in (L_1 \cap L_2)(\mathbb{R})$ with Fourier transform $\mathbb{F}u = \widehat{u} \in (L_1 \cap L_2)(\mathbb{R})$. Then for any $f \in L_1(\mathbb{R})$,*

$$\left(\mathbb{F}_u f\right)(x, \omega) = e^{-ix\omega} \left(\mathbb{F}_{u^{\star}}^{\#} \widehat{f}\right)(x, \omega), \tag{7.4.5}$$

for $(x, \omega) \in \mathbb{R}^2$, when $u^{\star}$ is defined by

$$u^{\star}(\xi) = \widehat{u}(-\xi), \text{ for } \xi \in \mathbb{R}.$$

Proof   The proof of (7.4.5) follows by applying (7.2.7) of Theorem 2 on p.331. Indeed, by considering

$$g(t) = \overline{u(t-x)}e^{i\omega t} = \left(M_\omega \, T_x \bar{u}\right)(t),$$

it follows from (7.1.9) and (7.1.11) of Theorem 2 on p.331 that

$$\widehat{g}(\xi) = \widehat{\bar{u}}(\xi - \omega)e^{-ix(\xi-\omega)} = \overline{\widehat{u}(\omega - \xi)}e^{-ix(\xi-\omega)}.$$

Hence, by (7.2.7) of Theorem 2 on p.331 and the definition of $\mathbb{F}_u$ in (7.4.1), we have, for fixed (x, ω),

$$
\begin{aligned}
\left(\mathbb{F}_u f\right)(x, \omega) &= \langle f, g \rangle = \frac{1}{2\pi}\langle \widehat{f}, \widehat{g} \rangle \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty} \widehat{f}(\xi)\widehat{u}(\omega - \xi)\, e^{ix(\xi-\omega)}\, d\xi \\
&= e^{-ix\omega}\frac{1}{2\pi}\int_{-\infty}^{\infty}\widehat{f}(\xi)u^\star(\xi - \omega)e^{ix\xi}d\xi \\
&= e^{-ix\omega}\left(\mathbb{F}_{u^\star}^{\#}\,\widehat{f}\right)(x, \omega).
\end{aligned}
$$

$\blacksquare$

   In view of the time-frequency localization identity (7.4.5), it is imperative to come up with window functions $u(x)$ such that both $xu(x) \in L_2(\mathbb{R})$ and $\omega\widehat{u}(\omega) \in L_2(\mathbb{R})$ in order to achieve finite window widths $2\Delta_u$ and $2\Delta_{u^\star}$, as defined in (7.4.3)–(7.4.4).

Example 1 The window function

$$u(x) = \chi_{(-\frac{1}{2}, \frac{1}{2})}(x)$$

with

$$\int_{-\infty}^{\infty} u(x)dx = \int_{-\frac{1}{2}}^{\frac{1}{2}} 1\,dx = 1$$

and center $x^\star = 0$ has finite Δ_u, but $\Delta_{u^\star} = \infty$.

Solution Clearly, with

$$\|u\|_2^2 = \int_{-\frac{1}{2}}^{\frac{1}{2}} 1^2 dx = 1,$$

we have

$$x^\star = \int_{-\infty}^{\infty} xu(x)^2 dx = \int_{-\frac{1}{2}}^{\frac{1}{2}} x\,dx = 0$$

and

$$\Delta_u = \int_{-\infty}^{\infty} x^2 u(x)^2 dx = \int_{-\frac{1}{2}}^{\frac{1}{2}} x^2 dx = \frac{1}{12}$$

is finite. On the other hand,

$$\widehat{u}(\omega) = \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{-i\omega x} dx = \frac{e^{-i\omega/2} - e^{i\omega/2}}{-i\omega}$$
$$= \frac{\sin(\omega/2)}{\omega/2},$$

and hence, $u^\star(\omega) = \widehat{u}(-\omega) = \widehat{u}(\omega)$ and

$$\int_{-\infty}^{\infty} |\omega\widehat{u}(\omega)|^2 \, d\omega = 4 \int_{-\infty}^{\infty} \sin^2(\frac{\omega}{2}) d\omega = \infty.$$
∎

The most commonly used time-window function is the Gaussian function defined by (7.1.13) on p.324.

To compute the window width, we differentiate both sides of (7.1.16) on p.325 with respect to α, and then set $\alpha = 1/(2\sigma^2)$, to yield

$$-\int_{-\infty}^{\infty} x^2 e^{-2(\frac{x}{2\sigma})^2} dx = -\frac{1}{2}\sqrt{\pi}\Big(\frac{1}{2\sigma^2}\Big)^{-3/2},$$

so that

$$\int_{-\infty}^{\infty} x^2 g_\sigma^2(x) dx = \frac{1}{(2\sigma\sqrt{\pi})^2} \frac{\sqrt{\pi}}{2} 2^{3/2}\sigma^3.$$

Since g_α is an even function, the center x^* of g_α is 0. On the other hand, since

$$\int_{-\infty}^{\infty} g_\sigma^2(x) dx = \frac{1}{(2\sigma\sqrt{\pi})^2} \int_{-\infty}^{\infty} e^{-2(\frac{x}{2\sigma})^2} dx$$
$$= \frac{1}{(2\sigma\sqrt{\pi})^2} (\sqrt{2}\sigma)\sqrt{\pi},$$

we have

$$\begin{aligned}
\big(\Delta_{g_\sigma}\big)^2 &= \int_{-\infty}^{\infty} x^2 g_\sigma^2(x) dx \Big/ \int_{-\infty}^{\infty} g_\sigma^2(x) dx \\
&= \frac{\sqrt{\pi}}{2} 2^{3/2}\sigma^3 \Big/ (\sqrt{2}\sigma\sqrt{\pi}) = \sigma^2,
\end{aligned}$$

so that the width of the window function g_σ is

$$2\Delta_{g_\sigma} = 2\sigma, \tag{7.4.6}$$

where σ is the standard deviation of the Gaussian function $g_\sigma(x)$.

To compute the window width of the Fourier transform $\widehat{g}_\sigma$ of g_σ, we re-write the Fourier transform $\widehat{g}_\alpha(\omega)$ in (7.1.20) on p.326 as

$$\widehat{g}_\sigma(\omega) = e^{-(\omega/2\eta)^2} \text{ with } \eta = \frac{1}{2\sigma}.$$

Then, we may conclude, by applying the result $\Delta_{g_\eta} = \eta$ in (7.4.6), that $\Delta_{\widehat{g}_\sigma} = \eta = \frac{1}{2\sigma}$; so that the width of the window function $\widehat{g}_\sigma$ is $2\Delta_{\widehat{g}_\sigma} = \frac{1}{\sigma}$. We summarize the above results in the following.

Theorem 2 | **Time-frequency window of Gaussian function** | *The radii of the window functions of Gaussian functions $g_\sigma(x)$ and $\widehat{g}_\sigma(\omega)$ are given by*

$$\Delta_{g_\sigma} = \sigma, \quad \Delta_{\widehat{g}_\sigma} = \frac{1}{2\sigma}; \tag{7.4.7}$$

and the area of the time-frequency localization window:

$$\left[-\Delta_{g_\sigma}, \ \Delta_{g_\sigma} \right] \times \left[-\Delta_{\widehat{g}_\sigma}, \ \Delta_{\widehat{g}_\sigma} \right] \tag{7.4.8}$$

in $\mathbb{R}^2$ is

$$\left(2\Delta_{g_\sigma}\right)\left(2\Delta_{\widehat{g}_\sigma}\right) = 2. \tag{7.4.9}$$

It turns out that the area $= 2$ is the smallest among all time-frequency localization windows $\left[-\Delta_u, \ \Delta_u \right] \times \left[-\Delta_{\widehat{u}}, \ \Delta_{\widehat{u}} \right]$, where $u \in (L_1 \cap L_2)(\mathbb{R})$, as asserted by the following so-called "uncertainty principle".

Theorem 3 | **Uncertainty principle** | *Let $u \in \left(L_1 \cap L_2\right)(\mathbb{R})$ with Fourier transform $\widehat{u}(\omega)$. Then*

$$\Delta_u \Delta_{\widehat{u}} \geq \frac{1}{2}, \tag{7.4.10}$$

where Δ_u or $\Delta_{\widehat{u}}$ may be infinite. Furthermore, equality in (7.4.10) holds if and only if

$$u(x) = cg_\sigma(x - b) \tag{7.4.11}$$

for $\sigma > 0, b \in \mathbb{R}$, and $c \neq 0$.

In other words, the Gaussian function is the only time-window function that provides optimal time-frequency localization. The proof of Theorem 3 is delayed to the end of this section.

After discussing the area of the time-frequency localization windows, we now return to study the localized (or short-time) Fourier transform $(\mathbb{F}_u f)(x, \omega)$. To recover $f(x)$ from $(\mathbb{F}_u f)(x, \omega)$, we have the following result.

Theorem 4 *Let $u \in (L_1 \cap L_2)(\mathbb{R})$ with Fourier transform $\widehat{u} \in (L_1 \cap L_2)(\mathbb{R})$ such that $u(0) \neq 0$. Then for any $f \in (L_1 \cap L_2)(\mathbb{R})$ with $\widehat{f} \in L_1(\mathbb{R})$,*

$$f(x) = \frac{1}{u(0)} \frac{1}{2\pi} \int_{-\infty}^{\infty} (\mathbb{F}_u f)(x, \omega) \, e^{ix\omega} d\omega. \tag{7.4.12}$$

The derivation of (7.4.12) follows from Theorem 1 by multiplying both sides of (7.4.5) with $e^{ix\omega}$ and then taking the integral; namely,

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} (\mathbb{F}_u f)(x, \omega) e^{ix\omega} d\omega = \frac{1}{2\pi} \int_{-\infty}^{\infty} (\mathbb{F}_{u\star}^{\#} \widehat{f})(x, \omega) d\omega$$

$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} \widehat{f}(\xi) e^{ix\xi} \Big(\frac{1}{2\pi} \int_{-\infty}^{\infty} \widehat{u}\big(-(\xi - \omega) \big) d\omega \Big) d\xi$$

$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} \widehat{f}(\xi) e^{ix\xi} \Big(\frac{1}{2\pi} \int_{-\infty}^{\infty} \widehat{u}(y) dy \Big) d\xi$$

$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} \widehat{f}(\xi) e^{ix\xi} \Big(\frac{1}{2\pi} \int_{-\infty}^{\infty} \widehat{u}(y) e^{i0y} dy \Big) d\xi$$

$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} \widehat{f}(\xi) e^{ix\xi} u(0) d\xi$$

$$= u(0) \Big(\frac{1}{2\pi} \int_{-\infty}^{\infty} \widehat{f}(\xi) e^{ix\xi} d\xi \Big) = u(0) f(x),$$

where Theorem 4 on p.332 is applied to both $\widehat{u}$ and $\widehat{f}$ which are assumed to be in $L_1(\mathbb{R})$.   ∎

Remark 3 If the Gaussian function $g_\sigma(x)$ defined in (7.1.13) on p.324 is used as the window function $u(x)$ in Theorem 4, then since g_σ and $\widehat{g}_\sigma$ are both in $(L_1 \cap L_2)(\mathbb{R})$ and since

$$g_\sigma(0) = \frac{1}{2\sqrt{\pi}\sigma},$$

it follows that the choice of $\sigma = 1/(2\sqrt{\pi})$ in (7.4.12) yields

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} (\mathbb{G}f)(x, \omega) e^{ix\omega} d\omega \tag{7.4.13}$$

for all $f \in L_1(\mathbb{R})$ with $\widehat{f} \in L_1(\mathbb{R})$, where $\mathbb{G}f$ is defined as follows.   ∎

Definition 4 $\boxed{\text{Gabor transform}}$ *The integral transform*

$$(\mathbb{G}f)(x, \omega) = \int_{-\infty}^{\infty} f(t) e^{-\pi(t-x)^2} e^{-it\omega} dt \tag{7.4.14}$$

of functions $f \in L_1(\mathbb{R})$ *is called the Gabor transform of* $f(x)$ *at the* (x, ω) *position of the time-frequency domain* $\mathbb{R}^2$.

Hence, every $f \in L_1(\mathbb{R})$ with Fourier transform $\widehat{f} \in L_1(\mathbb{R})$ can be recovered from its Gabor transform $(\mathbb{G}f)(x, \omega)$ by applying the formula (7.4.13). ∎

We end this section by providing the proof of Theorem 3.

Proof of Theorem 3   To prove this theorem, we may assume, without loss of generality, that the centers of $u(x)$ and $\widehat{u}(\omega)$ are $x^\star = 0$ and $\omega^\star = 0$, respectively (see Exercise 1). Hence,

$$\left(\Delta_u \Delta_{\widehat{u}} \right)^2 = \frac{1}{\|u\|_2^2 \, \|\widehat{u}\|_2^2} \left(\int_{-\infty}^{\infty} x^2 |u(x)|^2 dx \right) \left(\int_{-\infty}^{\infty} \omega^2 |\widehat{u}(\omega)|^2 d\omega \right). \quad (7.4.15)$$

In (7.4.15), we may apply Theorem 1 (iv) on p.320 and Parseval's identity on p.330 to conclude that

$$\int_{-\infty}^{\infty} \omega^2 |\widehat{u}(\omega)|^2 d\omega = \|\widehat{u'}\|_2^2 = 2\pi \|u'\|_2^2. \quad (7.4.16)$$

In addition, in the denominator of (7.4.15), we have $\|\widehat{u}\|_2^2 = 2\pi \|u\|_2^2$, again by Parseval's identity. Therefore, applying the Cauchy-Schwarz inequality, we obtain

$$\left(\Delta_u \Delta_{\widehat{u}} \right)^2 = \|u\|_2^{-4} \int_{-\infty}^{\infty} |xu(x)|^2 \int_{-\infty}^{\infty} |u'(x)|^2 dx$$

$$\geq \|u\|_2^{-4} \left(\int_{-\infty}^{\infty} |xu(x)\overline{u'(x)}| dx \right)^2 \quad (7.4.17)$$

$$\geq \|u\|_2^{-4} \left| \int_{-\infty}^{\infty} \text{Re} \, \{xu(x)\overline{u'(x)}\} dx \right|^2. \quad (7.4.18)$$

But since

$$x \frac{d}{dx} |u(x)|^2 = x \frac{d}{dx} u(x)\overline{u(x)}$$

$$= x \left(u(x)\overline{u'(x)} + u'(x)\overline{u(x)} \right)$$

$$= 2\text{Re} \, \{xu(x)\overline{u'(x)}\},$$

the right-hand side of (7.4.18) can be written as

$$\|u\|_2^{-4} \left\{ \frac{1}{2} \int_{-\infty}^{\infty} \left(x \frac{d}{dx} |u(x)|^2 \right) dx \right\}^2$$

$$= \frac{1}{4} \|u\|_2^{-4} \left\{ \left[x|u(x)|^2 \right]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} |u(x)|^2 dx \right\}^2$$

$$= \frac{1}{4} \|u\|_2^{-4} \, \|u\|_2^4 = \frac{1}{4}, \quad (7.4.19)$$

and hence, $\Delta_u \Delta_{\hat{u}} \geq \frac{1}{2}$. In both (7.4.16) and (7.4.19), we have assumed that $u \in PC(\mathbb{R})$. In addition, since $u \in L_2(\mathbb{R})$ or $|u|^2 \in L_1(\mathbb{R})$, the function $|u(x)|^2$ must decay to 0 faster than $\frac{1}{x}$, when $|x| \to \infty$. That (7.4.16) and (7.4.19) are valid for any $u \in L_1 \cap L_2(\mathbb{R})$ follows from a standard "density" argument of

$$\text{closure}_{L_2}\big(PC(\mathbb{R})\big) = L_2(\mathbb{R}).$$

Finally, if the inequality in (7.4.18) becomes equality, we recall from the theorem on the Cauchy-Schwarz inequality (see Theorem 3 on p.27) that

$$|xu(x)| = r|u'(x)| \tag{7.4.20}$$

for some constant $r > 0$ and

$$\pm \, \text{Re} \, xu(x)\overline{u'(x)} = |xu(x)u'(x)|. \tag{7.4.21}$$

From (7.4.20), we have

$$xu(x) = ru'(x)e^{i\theta(x)} \tag{7.4.22}$$

for some real-valued function $\theta(x)$. Hence, by (7.4.21) together with (7.4.22), we may conclude that

$$\pm\text{Re} \, r|u'(x)|^2 e^{i\theta(x)} = r|u'(x)|^2.$$

Thus $\pm\text{Re}(e^{i\theta(x)}) = 1$, which implies that $\pm e^{i\theta(x)}$ is the constant function 1. Therefore, (7.4.22) becomes

$$\frac{u'(x)}{u(x)} = \frac{1}{r}x \quad \text{or} \quad \frac{u'(x)}{u(x)} = \frac{-1}{r}x,$$

or equivalently,

$$u(x) = \tilde{c}e^{x^2/2r} \quad \text{or} \quad u(x) = \tilde{c}e^{-x^2/2r}.$$

But since $r > 0$ and $u(x) \in L_1(\mathbb{R})$, $u(x)$ cannot be $\tilde{c}e^{x^2/2r}$ and must be the Gaussian function

$$u(x) = \tilde{c}e^{-\alpha^2 x^2}$$

with $\alpha^2 = \frac{1}{2r} > 0$. In the above argument, we have assumed that the center $x^\star$ of the time-window function $u(x)$ is $x^\star = 0$. Therefore, in general, $u(x)$ can be formulated as

$$u(x) = cg_\sigma(x - b)$$

for $\sigma = \frac{1}{2\alpha}$ and some $c \neq 0$, $x^\star = b \in \mathbb{R}$.   ∎

Exercises

Exercise 1 Let $x^\star$ be the center of a localization window function $u(x)$ as defined by (7.4.3). Show that the center of the window function $\tilde{u}(x) = u(x + x^\star)$ is at the origin $x = 0$.

Exercise 2 Show that if $u \in L_2(\mathbb{R})$ and $xu(x) \in L_2(\mathbb{R})$, then $xu(x)^2 \in L_1(\mathbb{R})$.

Exercise 3 Show that $(\overline{\mathbb{F}\bar{u}})(\omega) = \hat{u}(-\omega)$.

Exercise 4 Let $m = 0, 1, 2, \ldots$. Compute $\int_{-\infty}^{\infty} x^m e^{-x^2} dx$.

Hint: Consider the cases of even and odd m, and for even m, refer to the calculation of $\int_{-\infty}^{\infty} x^2 e^{-\alpha x^2} dx$ for Δ_{g_σ}.

Exercise 5 Let $h(x) = e^{-ax^2}$, $a > 0$.

(a) Compute Δ_h.
(b) Compute $\Delta_{\hat{h}}$.
(c) Show that $\Delta_h \Delta_{\hat{h}} = \frac{1}{2}$.

Exercise 6 Let $g_c(x, t)$ be defined by

$$g_c(x, t) = \frac{1}{2\sqrt{ct\pi}} e^{-\frac{x^2}{4ct}},$$

where $c > 0$ is fixed and $t > 0$ is considered as the time variable.

(a) Compute $\dfrac{\partial}{\partial t} g_c(x, t)$.

(b) Compute $\dfrac{\partial}{\partial x} g_c(x, t)$.

(c) Compute $\dfrac{\partial^2}{\partial x^2} g_c(x, t)$.

(d) Derive the relationship between $\dfrac{\partial}{\partial t} g_c(x, t)$ and $\dfrac{\partial^2}{\partial x^2} g_c(x, t)$.

Exercise 7 Let $x^\star$ be the center of a localization window function $u(x)$, as introduced in Definition 3. Verify that

$$\int_{-\infty}^{\infty} (x - x^\star)|u(x)|^2 dx = 0.$$

Exercise 8 Let $h_1(x) = \chi_{[0, 1]}(x)$, and for $n = 2, 3, \ldots$, define $h_n(x)$ recursively by

$$h_n(x) = \left(h_{n-1} * h_1\right)(x) = \int_{-\infty}^{\infty} h_{n-1}(t - x)h_1(t)dt = \int_0^1 h_{n-1}(t - x)dt.$$

Prove by mathematical induction that the functions $H_n(x) = h_n(x + \frac{n}{2})$ are even, for all $n = 1, 2, \ldots$.

Hint: If $f(x)$ is even, then

$$\int_{-a}^{a} f(t - x)dt = \int_{-a}^{a} f(t + x)dt,$$

which can be derived by the change of variables of integration from t to $u = -t$.

Exercise 9 As a continuation of Exercise 8, determine the center of the localization window function $h_n(x)$ for each $n = 1, 2, \ldots$.

Exercise 10 Show that the time-frequency window for the localization window function $h_n(x)$ has a finite area for $n \geq 2$ by showing that

$$\triangle_{h_n} \triangle_{\widehat{h}_n} < \infty.$$

Exercise 11 (This is a very difficult problem and is included only for the very advanced students.) Let $a_n = \triangle_{h_n} \triangle_{\widehat{h}_n}$. Prove that

$$a_2 > a_3 > \cdots > a_n \to \frac{1}{2},$$

as $n \to \infty$. In other words, although $h_n(x)$ is not the Gaussian, it provides near-optimal time-frequency localization as governed by the uncertainty principle (in Theorem 3) for sufficiently large n.

7.5 Time-Frequency Bases

As already established in Theorem 1 on p.352 in the previous section, computation of the localized (or short-time) Fourier transform $(\mathbb{F}_u f)(x, \omega)$ of $f \in L_1(\mathbb{R})$ provides the quantity of the localized transform $(\mathbb{F}_{u^\star}^{\#} \widehat{f})(x, \omega)$ of the frequency content $\widehat{f}(\omega)$ with window function $u^\star(\xi) = \widehat{u}(-\xi)$. On the other hand, $f(x)$ can be recovered from $(\mathbb{F}_u f)(x, \omega)$ by the formula (7.4.12) on p.356. In particular, when the window function is the Gaussian function, then the localized Fourier transform is the Gabor transform $(\mathbb{G}f)(x, \omega)$, and every $f \in L_1(\mathbb{R})$ with Fourier transform $\widehat{f} \in L_1(\mathbb{R})$ can be recovered from its Gabor transform by applying the formula (7.4.13) on p.356.

In this section, we study the time-frequency analysis by considering sampling of the LFT (or STFT) $(\mathbb{F}_{\bar{u}} f)(x, \omega)$ of $f \in (L_1 \cap L_2)(\mathbb{R})$, where the complex conjugate $\overline{u(x)}$ of $u \in (L_1 \cap L_2)(\mathbb{R})$ is used as the window function. Let $\mathbb{Z}$ denote the set of all integers. By sampling $(\mathbb{F}_{\bar{u}} f)(x, \omega)$ at $(x, \omega) = (m, 2\pi k)$, with $m, k \in \mathbb{Z}$, we have

$$(\mathbb{F}_{\bar{u}} f)(m, 2\pi k) = \int_{-\infty}^{\infty} f(x)\overline{u(x - m)}\, e^{-i2\pi kx} dx$$

$$= \int_{-\infty}^{\infty} f(x)\overline{h_{m,k}(x)}dx = \langle f, h_{m,k} \rangle,$$

with $h_{m,k}(x)$ defined by

$$h_{m,k}(x) = u(x - m)\, e^{i2\pi kx}. \tag{7.5.1}$$

Remark 1 Since $e^{-i2\pi kx} = e^{-i2\pi k(x-m)}$, the functions $h_{m,k}(x)$ in (7.5.1) can be formulated as

$$h_{m,k}(x) = H_k(x - m), \tag{7.5.2}$$

with $H_k(x)$ defined by

$$H_k(x) = u(x)\, e^{i2\pi kx}.$$

Hence, while $H_k(x)$ localizes the frequency $k \in \mathbb{Z}$ of $f(x)$ only at the time sample point $m = 0$, $h_{m,k}(x)$ localizes the same frequency of $f(x)$ at any time instant $m \in \mathbb{Z}$.
∎

Remark 2 On the other hand, if the frequency $k \in \mathbb{Z}$ in the definition (7.5.1) is not an integer, such as

$$h_{m,kb} = u(x - m)\, e^{i2\pi bkx},$$

where $b \notin \mathbb{Z}$, then (7.5.2) does not apply. Later, we will also consider

$$(\mathbb{F}_{\bar{u}} f)(ma, 2\pi kb) = \langle f, h_{ma,kb} \rangle$$

with $a, b > 0$ and

$$h_{ma,kb}(x) = u(x - ma)\, e^{i2\pi kbx}. \tag{7.5.3}$$

∎

Example 1 Let $u(x)$ be the characteristic function of $[-\frac{1}{2}, \frac{1}{2})$; that is,

$$u(x) = \chi_{[-\frac{1}{2}, \frac{1}{2})}(x) = \begin{cases} 1, & \text{for } -\frac{1}{2} \le x < \frac{1}{2}, \\ 0, & \text{otherwise.} \end{cases}$$

Then the family $\{h_{m,k}(x)\}$, $m, k \in \mathbb{Z}$ defined by (7.5.1), constitutes an orthonormal basis of $L_2(\mathbb{R})$.

Solution For each $m \in \mathbb{Z}$,

$$\langle h_{m,k}, h_{m,\ell} \rangle = \int_{m-\frac{1}{2}}^{m+\frac{1}{2}} e^{i2\pi(k-\ell)x} dx = \delta_{k-\ell}.$$

For all $k, \ell \in \mathbb{Z}$ and $m \neq n$,

$$\langle h_{m,k}, h_{n,\ell} \rangle = 0,$$

since the supports of $h_{m,k}$ and $h_{n,\ell}$ do not overlap. Hence,

$$\langle h_{m,k}, h_{n,\ell} \rangle = \delta_{m-n}\delta_{k-\ell};$$

or $\{h_{m,k}(x)\}$, $m, k \in \mathbb{Z}$, is an orthonormal family.

In addition, for any $f \in L_2(\mathbb{R})$, by setting $f_m(x) = f(x+m)$, where $-\frac{1}{2} \leq x \leq \frac{1}{2}$, namely:

$$f_m(x) = u(x)f(x+m) = \chi_{[-\frac{1}{2},\frac{1}{2})}(x)f(x+m), \quad -\frac{1}{2} \leq x \leq \frac{1}{2},$$

and extending it to $\mathbb{R}$ periodically such that $f_m(x+1) = f_m(x)$, then for each fixed $m \in \mathbb{Z}$, the Fourier series

$$(Sf_m)(x) \ = \ \sum_{k=-\infty}^{\infty} a_k(f_m)\, e^{i2\pi kx}$$

of $f_m(x)$ converges to $f_m(x)$ in $L_2[-\frac{1}{2}, \frac{1}{2}]$, where

$$\begin{aligned}
a_k(f_m) \ &= \ \int_{-\frac{1}{2}}^{\frac{1}{2}} f_m(t)\, e^{-i2\pi kt}\, dt = \int_{-\frac{1}{2}}^{\frac{1}{2}} f(t+m)\, e^{-i2\pi kt}\, dt \\
&= \int_{m-\frac{1}{2}}^{m+\frac{1}{2}} f(y)\, e^{-i2\pi k(y-m)}\, dy \\
&= \int_{-\infty}^{\infty} f(t)u(t-m)\, e^{-i2\pi kt}\, dt = \langle f, h_{m,k}\rangle.
\end{aligned}$$

Thus, for each $m \in \mathbb{Z}$, we have

$$f(x)u(x-m) = f_m(x-m)u(x-m) = \big((Sf_m)(x-m)\big)u(x-m)$$

$$= \sum_{k=-\infty}^{\infty} a_k(f_m)\, e^{i2\pi k(x-m)}u(x-m) = \sum_{k=-\infty}^{\infty} \langle f, h_{m,k}\rangle h_{m,k}(x).$$

Hence, summing both sides over all $m \in \mathbb{Z}$ yields

$$\begin{aligned}
f(x) \ &= \ \sum_{m=-\infty}^{\infty} f(x)\chi_{[m-\frac{1}{2},m+\frac{1}{2})}(x) = \sum_{m=-\infty}^{\infty} f(x)\chi_{[-\frac{1}{2},\frac{1}{2})}(x-m) \\
&= \sum_{m=-\infty}^{\infty} f(x)u(x-m) = \sum_{m=-\infty}^{\infty}\sum_{k=-\infty}^{\infty} \langle f, h_{m,k}\rangle h_{m,k}(x),
\end{aligned}$$

where the convergence is in the $L_2(\mathbb{R})$-norm (see Exercise 1).   ∎

Remark 3 The limitation of the window function $u(x) = \chi_{[-\frac{1}{2}, \frac{1}{2})}(x)$ in the above example is that it provides very poor frequency localization, as discussed in Example 1 on p.353. Unfortunately, the formulation of $h_{m,k}(x)$ in (7.5.1) cannot be improved by too much, as dictated by the so-called "Balian-Low" restriction to be stated in Theorem 2 later in this section (see also Remark 5).

∎

Definition 1 $\boxed{\text{Frame}}$ *A family of functions* $\{h_\alpha(x)\}$ *in* $L_2(\mathbb{R})$, $\alpha \in J$, *where J is some infinite index set, such as* $\mathbb{Z}$ *and* $\mathbb{Z}^2$, *is called a frame of* $L_2(\mathbb{R})$, *if there exist constants A and B, with* $0 < A \le B < \infty$, *such that*

$$A\|f\|_2^2 \le \sum_{\alpha \in J} |\langle f, h_\alpha \rangle|^2 \le B\|f\|_2^2, \qquad (7.5.4)$$

for all $f \in L_2(\mathbb{R})$. *Here, A and B are called frame bounds. Furthermore, if A and B can be so chosen that* $A = B$ *in* (7.5.4), *then* $\{h_\alpha\}$ *is called a tight frame.*

Remark 4 If $\{h_\alpha\}$, $\alpha \in J$, is a tight frame with frame bound $A = B$, then the family $\{\widetilde{h}_\alpha(x)\}$, defined by

$$\widetilde{h}_\alpha(x) = \frac{1}{\sqrt{A}} h_\alpha(x), \quad \alpha \in J,$$

satisfies Parseval's identity

$$\|f\|_2^2 = \sum_{\alpha \in J} |\langle f, \widetilde{h}_\alpha \rangle|^2, \quad f \in L_2(\mathbb{R}). \qquad (7.5.5)$$

∎

Recall that an orthonormal basis of $L_2(\mathbb{R})$ also satisfies Parseval's identity (7.5.5). To understand the identity (7.5.5) for the tight frame $\{h_\alpha\}$, let us consider the function $f(x) = h_{\alpha_0}(x)$ for any fixed index $\alpha_0 \in J$, so that

$$\|\widetilde{h}_{\alpha_0}\|_2^2 = \sum_{\alpha \in J} |\langle \widetilde{h}_{\alpha_0}, \widetilde{h}_\alpha \rangle|^2$$

$$= \|\widetilde{h}_{\alpha_0}\|_2^4 + \sum_{\alpha \ne \alpha_0} |\langle \widetilde{h}_{\alpha_0}, \widetilde{h}_\alpha \rangle|^2$$

or

$$\|\widetilde{h}_{\alpha_0}\|_2^2 \left(1 - \|\widetilde{h}_{\alpha_0}\|_2^2\right) = \sum_{\alpha \ne \alpha_0} |\langle \widetilde{h}_{\alpha_0}, \widetilde{h}_\alpha \rangle|^2.$$

Since the right-hand side is non-negative, we see that

$$\|\widetilde{h}_{\alpha_0}\|_2^2 \leq 1.$$

In addition, if $\|\widetilde{h}_{\alpha_0}\|_2 = 1$, then the right-hand side vanishes, or $\langle \widetilde{h}_{\alpha_0}, h_\alpha \rangle = 0$ for all $\alpha \neq \alpha_0$. Thus if $\|\widetilde{h}_\alpha\|_2 = 1$ for each $\alpha \in J$, then $\{\widetilde{h}_\alpha\}_{\alpha \in J}$ is an orthonormal family.

Furthermore, observe that any frame $\{h_\alpha\}$ of $L_2(\mathbb{R})$ is complete in $L_2(\mathbb{R})$, meaning that span$\{h_\alpha : \alpha \in J\}$ is "dense" in $L_2(\mathbb{R})$ (more precisely, the completion of span$\{h_\alpha : \alpha \in J\}$ is $L_2(\mathbb{R})$). Indeed, if $\{h_\alpha\}$ is not complete in $L_2(\mathbb{R})$, then there would exist some non-trivial $f \in L_2(\mathbb{R})$ which is orthogonal to each h_α, violating the lower-bound frame condition, in that

$$0 < A\|f\|_2^2 \leq \sum_{\alpha \in J} |\langle f, h_\alpha \rangle|^2 = 0.$$

Let us summarize the above derivations in the following theorem.

Theorem 1 *Let $\{h_\alpha\}$, $\alpha \in J$, be a tight frame of $L_2(\mathbb{R})$ with frame bound $A = B = 1$. Then*

(a) $\|h_\alpha\|_2 \leq 1$ *for all $\alpha \in J$;*
(b) *the $L_2(\mathbb{R})$ closure of* span$\{h_\alpha : \alpha \in J\}$ *is the entire $L_2(\mathbb{R})$ space;*
(c) *if $\|h_\alpha\|_2 = 1$ for all $\alpha \in J$, then $\{h_\alpha\}$ is an orthonormal basis of $L_2(\mathbb{R})$.*

We now return to the study of time-frequency analysis by stating the Balian-Low restriction as in Theorem 2 below and a sampling criterion for achieving good time-frequency localization, in Theorem 3(c) later in this section.

Theorem 2 $\boxed{\textbf{Balian-Low restriction}}$ *Let $\{h_{m,k}(x)\}$, $(m, k) \in \mathbb{Z}^2$, be defined by (7.5.1) with window function $u(x) \in (L_1 \cap L_2)(\mathbb{R})$. Then a necessary condition for $\{h_{m,k}(x)\}$ to be a frame of $L_2(\mathbb{R})$ is that at least one of*

$$\int_{-\infty}^{\infty} |xu(x)|^2 dx \quad \text{and} \quad \int_{-\infty}^{\infty} |\omega\widehat{u}(\omega)|^2 d\omega$$

is equal to ∞.

Remark 5 Recall from Theorem 1 (iv) on p.320 and Theorem 1 on p.330 that

$$\int_{-\infty}^{\infty} |\omega\widehat{f}(\omega)|^2 d\omega = \int_{-\infty}^{\infty} |(\mathbb{F}f')(\omega)|^2 d\omega$$
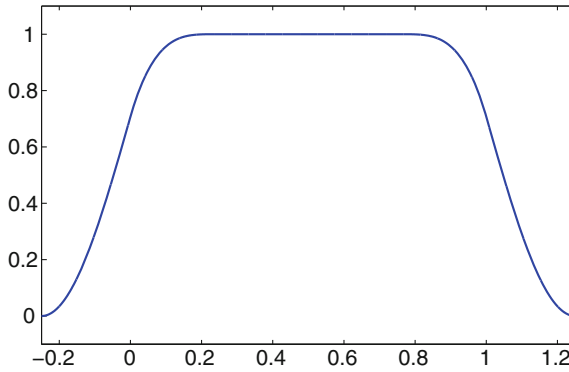
$$= 2\pi \int_{-\infty}^{\infty} |f'(x)|^2 dx.$$

Hence, if $u(x)$ is the window function in $\{h_{m,k}(x)\}$ with finite window width (that is, $\Delta_u < \infty$), then for $\{h_{m,k}(x)\}$ to be a frame of $L_2(\mathbb{R})$, it is necessary that

$$\int_{-\infty}^{\infty} |u'(x)|^2 dx = \infty$$

according to the Balian-Low restriction in Theorem 2. Consequently, any compactly supported smooth function, such as any B-spline $H_n(x) = h_n(x + \frac{n}{2})$ for $n \geq 2$ (defined in Exercise 8 of the previous section), cannot be used as the window function $u(x)$ in (7.5.1) to achieve good time-frequency localization. ∎

On the other hand, we have the following result.

Theorem 3 *Let $\{h_{ma,kb}(x)\}$, $(m, k) \in \mathbb{Z}^2$, be defined by (7.5.3) with window function $u \in (L_1 \cap L_2)(\mathbb{R})$, where $a, b > 0$.*

(a) *For $ab > 1$, there does not exist any window function $u(x)$ for which $\{h_{ma,kb}(x)\}$, $(m, k) \in \mathbb{Z}^2$, is complete in $L_2(\mathbb{R})$.*
(b) *For $ab = 1$ (such as $a = 1$ and $b = 1$ in (7.5.1)), there exists $u(x) \in (L_1 \cap L_2)(\mathbb{R})$ such that $\{h_{m,k}(x)\}$ is a frame (such as an orthonormal basis in Example 1), but the time-frequency window*

$$[-\Delta_u, \Delta_u] \times [-\Delta_{\widehat{u}}, \Delta_{\widehat{u}}]$$

necessarily has infinite area, namely:

$$\Delta_u \Delta_{\widehat{u}} = \infty.$$

(c) *For $0 < ab < 1$, there exists $u \in (L_1 \cap L_2)(\mathbb{R})$ such that $\Delta_u \Delta_{\widehat{u}} < \infty$ and the corresponding family $\{h_{ma,kb}(x)\}$, $(m, k) \in \mathbb{Z}^2$, is a tight frame of $L_2(\mathbb{R})$.*

Remark 6 Let a and b in (7.5.3) be restricted by $0 < ab < 1$ to achieve good time-frequency localization, as guaranteed by Theorem 1(c). Then the frame $\{h_{ma,kb}(x)\}$ of $L_2(\mathbb{R})$ in (7.5.3), with window function $u(x)$ (that satisfies $\Delta_u \Delta_{\widehat{u}} < \infty$) cannot be formulated as the translation (by mb, $m \in \mathbb{Z}$) of some localized function

$$H_{kb}(x) = u(x) \, e^{i2\pi kbx}$$

as in (7.5.2) for $h_{m,k}(x)$, where $a = b = 1$. Consequently, the computational aspect of time-frequency analysis is much less effective. ∎

The good news is that by replacing $e^{i2\pi kx}$ with the sine and/or cosine functions to formulate a frequency basis, such as

$$c_k(x) = \sqrt{2} \cos(k + \frac{1}{2})\pi x, \quad k = 0, 1, \dots \tag{7.5.6}$$

(that is, the family W_3 introduced in (6.2.17) of Sect. 6.2 on p.284 with $d = 1$, as the orthonormal basis of $L_2[0, 1]$ for the Fourier cosine series of type II), then localization by window functions $u(x)$ with finite time-frequency windows (that is,

$\Delta_u \Delta_{\widehat{u}} < \infty$) is feasible by formulation of the local basis functions $h^c_{mn}(x)$ as integer translates, such as

$$h^c_{m,k}(x) = H^c_k(x - m) = u(x - m)\, c_k(x - m), \quad m, k \in \mathbb{Z},$$

of the localized frequency basis

$$H^c_k(x) = u(x)\, c_k(x), \; k \in \mathbb{Z}$$

(compare this with (7.5.2)). Recall from Theorem 5 on p.284 that $c_k(x), \, k = 0, 1, \ldots$ defined by (7.5.6) constitute an orthonormal basis of $L_2[0, 1]$.

In this chapter, we will only focus on those computational efficient window functions $u(x)$ that satisfy the admissible conditions to be defined as follows:

Definition 2 $\boxed{\textbf{Admissible window function}}$ *A function $u(x)$ is said to be an admissible window function, if it satisfies the following conditions*:

(i) *there exists some positive number δ, with $0 < \delta < \frac{1}{2}$, such that*

$$u(x) = 1, \; for \; x \in [\delta, 1 - \delta],$$

and

$$u(x) = 0, \; for \; x \notin [-\delta, 1 + \delta];$$

(ii) $0 \leq u(x) \leq 1$;

(iii) *$u(x)$ is symmetric with respect to $x = \frac{1}{2}$, namely*:

$$u\left(\frac{1}{2} - x\right) = u\left(\frac{1}{2} + x\right), \; for \; all \; x;$$

(iv) *both $u(x)$ and $u'(x)$ are piecewise continuous on $[-\delta, 1 + \delta]$; and*

(v) *$u^2(x) + u^2(-x) = 1$, for $x \in [-\delta, \delta]$.*

(See Fig. 7.2 for the graphical display of a typical admissible window function $u(x)$ to be discussed in Example 2.)

Remark 7 It follows from (i), (ii) and (iv) that

$$\int_{-\infty}^{\infty} x^2\, u^2(x)dx < \infty \; \text{ and } \; \int_{-\infty}^{\infty} |u'(x)|^2 dx < \infty,$$

so that by Theorem 1 (iv) on p.320 and Parseval's theorem on p.330,

$$\int_{-\infty}^{\infty} \omega^2 |\widehat{u}(\omega)|^2 d\omega = \int_{-\infty}^{\infty} \left|(\mathbb{F}u')\right|^2 d\omega = \frac{1}{2\pi}\int_{-\infty}^{\infty} |u'(x)|^2 dx < \infty.$$

Fig. 7.2. Admissible window function $u(x)$ defined in (7.5.8)

Hence, $u(x)$ provides good time-frequency localization, meaning that $\Delta_u \Delta_{\widehat{u}} < \infty$. Furthermore, conditions (i), (iii) and (v) assure that

$$\sum_{m=-\infty}^{\infty} u^2(x - m) = 1, \text{ for all} x \in \mathbb{R} \qquad (7.5.7)$$

(see Exercise 4)   ∎

Example 2 Let $0 < \delta < \frac{1}{2}$ and $u(x)$ be the function defined by

$$u(x) = \begin{cases} 0, & x < -\delta \text{ or } x > 1 + \delta; \\ \dfrac{1}{\sqrt{2}}\left(1 + \sin \dfrac{\pi}{2\delta}x\right), & -\delta \leq x < 0; \\ \sqrt{1 - \dfrac{1}{2}\left(1 - \sin \dfrac{\pi}{2\delta}x\right)^2}, & 0 \leq x < \delta; \\ 1, & \delta \leq x \leq 1 - \delta; \\ \sqrt{1 - \dfrac{1}{2}\left(1 - \sin \dfrac{\pi}{2\delta}(1 - x)\right)^2}, & 1 - \delta < x \leq 1; \\ \dfrac{1}{\sqrt{2}}\left(1 + \sin \dfrac{\pi}{2\delta}(1 - x)\right), & 1 < x \leq 1 + \delta, \end{cases} \qquad (7.5.8)$$

with the graph of $y = u(x)$ shown in Fig. 7.2. Verify that $u(x)$ is an admissible window function.

Solution The verification of (i), (ii), (iii) and (v) for $u(x)$ is straightforward. Here we only verify that $u'(x)$ exists for any $x \in \mathbb{R}$, which is reduced to the points $x_0 = -\delta, 0, \delta, 1 - \delta, 1, 1 + \delta$ since it is clear that $u'(x)$ exists elsewhere. Since $u(x)$ is continuous (see Exercise 6), to verify that $u'(x_0)$ exists, it is enough to show that the left-hand and right-hand limits of $u'(x)$ at x_0 exist and are equal.

For $x_0 = -\delta$, clearly

$$\lim_{x \to -\delta^-} u'(x) = \lim_{x \to -\delta^-} 0 = 0.$$

On the other hand,

$$\lim_{x \to -\delta^+} u'(x) = \lim_{x \to -\delta^+} \frac{1}{\sqrt{2}} \cos(\frac{\pi}{2\delta}x) \frac{\pi}{2\delta} = \frac{1}{\sqrt{2}} \frac{\pi}{2\delta} \cos(-\frac{\pi}{2}) = 0.$$

Thus $u'(-\delta)$ exists.

For $x_0 = 0$,

$$\lim_{x \to 0^-} u'(x) = \lim_{x \to 0^-} \frac{1}{\sqrt{2}} \cos(\frac{\pi}{2\delta}x) \frac{\pi}{2\delta} = \frac{1}{\sqrt{2}} \frac{\pi}{2\delta} \cos 0 = \frac{\pi}{2\sqrt{2}\delta};$$

$$\lim_{x \to 0^+} u'(x) = \lim_{x \to 0^+} \frac{\pi}{4\delta} \Big(1 - \frac{1}{2}\big(1 - \sin\frac{\pi}{2\delta}x\big)^2\Big)^{-\frac{1}{2}} \big(1 - \sin\frac{\pi}{2\delta}x\big) \cos\frac{\pi}{2\delta}x$$

$$= \frac{\pi}{4\delta} \Big(\frac{1}{2}\Big)^{-\frac{1}{2}} = \frac{\pi}{2\sqrt{2}\delta}.$$

Therefore $u'(0)$ exists.

The verification of the existence of $u'(x)$ at $x_0 = \delta$ is left as an exercise (see Exercise 7). The existence of $u'(x)$ at $x_0 = 1-\delta, 1, 1+\delta$ follows from the symmetry of $u(x)$. ∎

Theorem 4 | **Malvar wavelets** | *Let $u(x)$ be an admissible window function that satisfies the conditions* (i)–(v) *in Definition 2. Then $u(x)$ has the property*

$$\Delta_u \Delta_{\widehat{u}} < \infty, \tag{7.5.9}$$

and the family $\{\psi_{m,k}(x)\}$ of functions defined by

$$\psi_{m,k}(x) = u(x - m)c_k(x - m)$$
$$= \sqrt{2}u(x - m)\cos\big((k + \frac{1}{2})\pi(x - m)\big), \; m \in \mathbb{Z}, k \geq 0, \tag{7.5.10}$$

where $c_k(x)$ is defined in (7.5.6), *constitutes an orthonormal basis of $L_2(\mathbb{R})$.*

Remark 8   The functions $\psi_{m,k}(x)$ are often called "Malvar wavelets" in the literature. For the discrete time-frequency setting, with $c_k(x)$ in (7.5.6) replaced by the DCT, the discrete formulation of $\psi_{m,k}(x)$ in (7.5.10) provides an enhanced version of block DCT since the window function $u(x)$ "smooths" the boundary of the tiles and therefore reduces "blocky artifacts", particularly for 8×8 tiles for the JPEG image compression standard as studied in Sect. 5.4 of Chap. 5. ∎
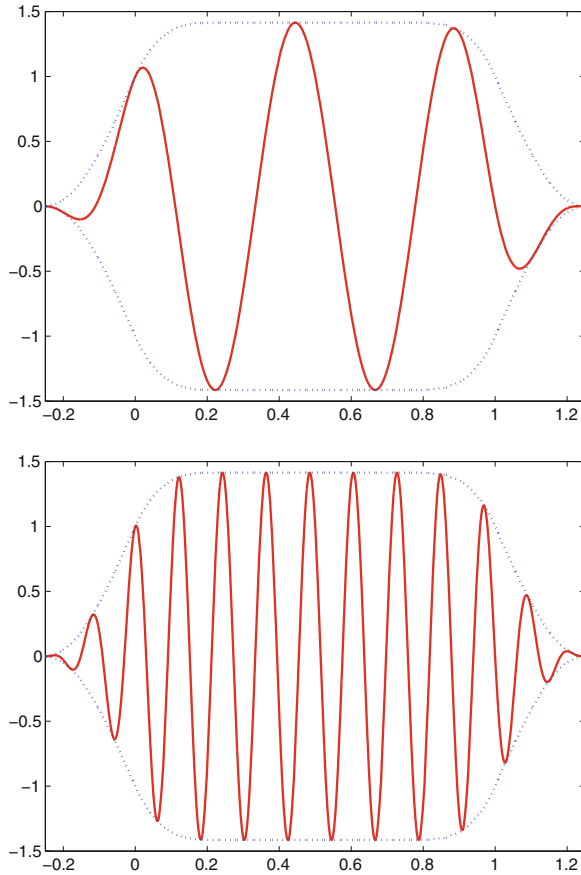
Fig. 7.3. Malvar wavelets $\psi_{0,4}$ (*top*) and $\psi_{0,16}$ (*bottom*). The envelope of $\cos(k+\frac{1}{2})\pi x$ is the dotted graph of $y = \pm\sqrt{2}u(x)$

Let $u(x)$ be the admissible window function given by (7.5.8) with $\delta = \frac{1}{4}$. We display the graphs of $u(x)$, $\psi_{0,4}$ and $\psi_{0,16}$ in Fig. 7.3.

Proof of Theorem 4   By Theorem 1 or Theorem 4 on p.46, to show that $\psi_{m,k}$, $m \in \mathbb{Z}$, $k \geq 0$ constitute an orthonormal basis for $L_2(\mathbb{R})$, it is sufficient to show $\|\psi_{m,k}\|_2 = 1$ and that Parseval's identity

$$\|f\|_2^2 = \sum_{m\in\mathbb{Z}}\sum_{k=0}^{\infty} |\langle f, \psi_{m,k}\rangle|^2, \text{ for all } f \in L_2(\mathbb{R}), \qquad (7.5.11)$$

holds.

Observe that by the symmetry of $u(x)$ and the property of the cosine function, we have

$$u(1-x) = u(x), \ u(2-x) = u(x-1),$$
$$c_k(-x) = c_k(x), \ c_k(2-x) = -c_k(x).$$

In addition, by property (v) of $u(x)$ and symmetry of $u(x)$, we also have

$$u^2(x) + u^2(2-x) = 1, \ \text{for} |x-1| \le \delta \tag{7.5.12}$$

(see Exercise 5). Thus,

$$\begin{aligned}
\|\psi_{m,k}\|_2^2 &= \langle \psi_{m,k}, \psi_{m,k} \rangle = \langle \psi_{0,\ell}, \psi_{0,k} \rangle \\
&= \int_{-\delta}^{1+\delta} c_k^2(x) u^2(x) dx \\
&= \left(\int_{-\delta}^{0} + \int_{0}^{1} + \int_{1}^{1+\delta} \right) c_k^2(x) u^2(x) dx \\
&= \int_{0}^{\delta} c_k^2(-y) u^2(-y) dy + \int_{0}^{1} c_k^2(x) u^2(x) dx + \int_{1-\delta}^{1} c_k^2(2-z) u^2(2-z) dz \\
&= \int_{0}^{\delta} c_k^2(x) \left(u^2(-x) + u^2(x) \right) dx + \int_{\delta}^{1-\delta} c_k^2(x) u^2(x) dx \\
&\quad + \int_{1-\delta}^{1} c_k^2(x) \left(u^2(x) + u^2(2-x) \right) dx \\
&= \int_{0}^{1} c_k^2(x) \cdot 1 dx = 1,
\end{aligned}$$

where the second equality follows from properties (i) and (v) of $u(x)$ and (7.5.12), and the fifth equality is a matter of change of variables of integration (with $y = -x$ and $z = 2 - x$).

Next let us verify Parseval's identity (7.5.11). In this regard, for $f \in L_2(\mathbb{R})$, set

$$f_m(x) = f(x+m) u(x),$$

where $m \in \mathbb{Z}$. Then we have

$$\begin{aligned}
\langle f, \psi_{m,k} \rangle &= \int_{-\infty}^{\infty} f(x) u(x-m) c_k(x-m) dx \\
&= \int_{-\infty}^{\infty} f(x+m) u(x) c_k(x) dx = \int_{-\delta}^{1+\delta} f_m(x) c_k(x) dx \\
&= \left(\int_{-\delta}^{0} + \int_{0}^{1} + \int_{1}^{1+\delta} \right) f_m(x) c_k(x) dx
\end{aligned}$$

$$= \int_0^\delta f_m(-x)c_k(-x)dx + \int_0^1 f_m(x)c_k(x)dx + \int_{1-\delta}^1 f_m(2-x)c_k(2-x)dx$$

$$= \int_0^1 \Big(f_m(-x)\chi_{[0,\delta]}(x) + f_m(x) - f_m(2-x)\chi_{[1-\delta,1]}(x)\Big)c_k(x)dx,$$

where the fact $c_k(-x) = c_k(x)$, $c_k(2-x) = -c_k(x)$ has been used. Since $\{c_k(x) : k = 0, 1, \ldots\}$ is an orthonormal basis of $L_2[0,1]$, by Parseval's identity applied to this basis, we have

$$\sum_{k=0}^\infty |\langle f, \psi_{m,k}\rangle|^2$$

$$= \int_0^1 \Big| f_m(-x)\chi_{[0,\delta]}(x) + f_m(x) - f_m(2-x)\chi_{[1-\delta,1]}(x)\Big|^2 dx$$

$$= \int_0^1 \Big(|f_m(-x)|^2\chi_{[0,\delta]}(x) + 2\mathrm{Re}\big(f_m(-x)\overline{f_m(x)}\big)\chi_{[0,\delta]}(x) + |f_m(x)|^2$$

$$\qquad -2\mathrm{Re}\big(f_m(x)\overline{f_m(2-x)}\big)\chi_{[1-\delta,1]}(x) + |f_m(2-x)|^2\chi_{[1-\delta,1]}(x)\Big)dx$$

$$= \int_0^\delta |f_m(-x)|^2 dx + \mathrm{I}_m + \int_0^1 |f_m(x)|^2 dx - \mathrm{II}_m + \int_{1-\delta}^1 |f_m(2-x)|^2 dx$$

$$= \int_{-\delta}^0 |f_m(y)|^2 dx + \mathrm{I}_m + \int_0^1 |f_m(x)|^2 dx - \mathrm{II}_m + \int_1^{1+\delta} |f_m(y)|^2 dy$$

$$= \int_{-\delta}^{1+\delta} |f_m(x)|^2 dx + \mathrm{I}_m - \mathrm{II}_m$$

$$= \int_{-\infty}^\infty |f(x)u(x-m)|^2 dx + \mathrm{I}_m - \mathrm{II}_m,$$

where

$$\mathrm{I}_m = \int_0^\delta 2\mathrm{Re}\big(f_m(-x)\overline{f_m(x)}\big)dx;$$

$$\mathrm{II}_m = \int_{1-\delta}^1 2\mathrm{Re}\big(f_m(x)\overline{f_m(2-x)}\big)dx.$$

Observe that

$$\mathrm{I}_{m+1} = \int_0^\delta 2\mathrm{Re}\big(f(m+1-x)\overline{f(m+1+x)}u(-x)u(x)\big)dx$$

$$= \int_{1-\delta}^1 2\mathrm{Re}\big(f(m+y)\overline{f(m+2-y)}u(y-1)u(1-y)\big)dy$$

$$= \int_{1-\delta}^1 2\mathrm{Re}\big(f(m+y)\overline{f(m+2-y)}u(2-y)u(y)\big)dy = \mathrm{II}_m.$$

Hence, it follows that

$$\sum_{m=-M}^{N}\sum_{k=0}^{\infty}|\langle f, \psi_{m,k}\rangle|^2 = \sum_{m=-M}^{N}\int_{-\infty}^{\infty}|f(x)u(x-m)|^2 dx + \mathrm{I}_{-M} - \mathrm{II}_N. \quad (7.5.13)$$

Since $f \in L_2(\mathbb{R})$, $\mathrm{I}_{-M} \to 0$, $\mathrm{II}_N \to 0$ as $M, N \to \infty$. In addition, by (7.5.7),

$$\lim_{M,N\to\infty}\sum_{m=-M}^{N}\int_{-\infty}^{\infty}|f(x)u(x-m)|^2 dx$$

$$= \int_{-\infty}^{\infty}|f(x)|^2\sum_{m=-\infty}^{\infty}u^2(x-m)dx = \int_{-\infty}^{\infty}|f(x)|^2 dx.$$

Therefore, by taking limit with $M, N \to \infty$ on both sides of (7.5.13), we obtain
(7.5.11). ∎

<div align="center">

Exercises

</div>

Exercise 1 Provide the details in showing that

$$\|\sum_{m=-N}^{N}f(x)u(x-m) - f(x)\|_2 \to 0$$

in Example 1.

Exercise 2 Let $g_{m,k}(x) = u(x - \frac{m}{2})e^{i2\pi kx}$, where $u(x)$ is the characteristic function of $[-\frac{1}{2}, \frac{1}{2})$. Hence, $g_{2m,k}(x) = h_{m,k}(x)$ in Example 1. Show that the family $\{g_{m,k}(x)\}$, $m, k \in \mathbb{Z}$, is a frame of $L_2(\mathbb{R})$.

    Hint: For the upper frame bound, consider the two sub-families $g_{2\ell,k}(x)$ and
$g_{2\ell+1,k}(x) = u(x - \frac{2\ell+1}{2})e^{i2\pi kx} = h_{\ell,k}(x - \frac{1}{2})$, $\ell \in \mathbb{Z}$, separately.

Exercise 3 If $\{h_\alpha(x)\}$, $\alpha \in J$, is a tight frame of $L_2(\mathbb{R})$ with frame bound $A = B$, justify that the family $\{\widetilde{h}_\alpha(x)\}$ introduced in Remark 4 is a tight frame with frame bound $= 1$.

Exercise 4 Let $u(x)$ be an admissible window function in Definition 2. Show that $u(x)$ satisfies the property (7.5.7) in that $\{u^2(x-m)\}$, $m \in \mathbb{Z}$, provides a partition of unity.

Exercise 5 Show that an admissible window function $u(x)$ satisfies (7.5.12).

Exercise 6 Let $u(x)$ be the function defined by (7.5.8). Show that $u(x)$ is continuous on $\mathbb{R}$.

Exercise 7 Let $u(x)$ be the function defined by (7.5.8). Verify that $u'(\delta)$ exists.

7.6 Appendix on Integration Theory

The study of Fourier transform and time-frequency analysis in this chapter requires some basic results from the integration theory. For completeness, we include, in this Appendix, a brief discussion of only those theorems that are needed for our study in this book. Of course, we only present these results for piecewise continuous functions to avoid the need of Lebesgue integration and more advanced theory. The reader is referred to more advanced textbooks in Real Analysis and Integration Theory.

Theorem 1 $\boxed{\textbf{Fubini's theorem}}$ *Let J_1, J_2 be two finite or infinite intervals and $f(x, y)$ be a function defined on $J_1 \times J_2$, such that for each $x \in J_1$, $g(y) = f(x, y)$ is in $PC(J_2)$ and for each $y \in J_2$, $h(x) = f(x, y)$ is in $PC(J_1)$. Suppose that either $f(x, y) \geq 0$ on $J_1 \times J_2$, or at least one of the two integrals*

$$\int_{J_1} \left(\int_{J_2} |f(x, y)| dy \right) dx, \quad \int_{J_2} \left(\int_{J_1} |f(x, y)| dx \right) dy,$$

is finite. Then the order of integration over J_1 and J_2 can be interchanged; that is,

$$\int_{J_1} \left(\int_{J_2} f(x, y) dy \right) dx = \int_{J_2} \left(\int_{J_1} f(x, y) dx \right) dy.$$

∎

Theorem 2 $\boxed{\textbf{Minkowski's integral inequality}}$ *Let J_1, J_2 be two finite or infinite intervals and $f(x, y)$ be a function defined on $J_1 \times J_2$, such that for each $x \in J_1$, $g(y) = f(x, y)$ is in $PC(J_2)$ and for each $y \in J_2$, $h(x) = f(x, y)$ is in $PC(J_1)$. Then for any $1 \leq p < \infty$,*

$$\left(\int_{J_1} \left| \int_{J_2} f(x, y) dy \right|^p dx \right)^{\frac{1}{p}} \leq \int_{J_2} \left(\int_{J_1} |f(x, y)|^p dx \right)^{\frac{1}{p}} dy. \tag{7.6.1}$$

The integrals in (7.6.1) are allowed to be infinite. The inequality (7.6.1) can be obtained from the Minkowski inequality for sequences. Indeed, for $f \in C(J_1 \times J_2)$, where $J_1 = [a, b]$ and $J_2 = [c, d]$ are bounded intervals, we may consider any partitions of J_1 and J_2, namely:

$$a = x_0 < x_1 < \cdots < x_m = b; \quad c = y_0 < y_1 < \cdots < y_n = d.$$

Set $\triangle x_j = x_j - x_{j-1}$, $\triangle y_k = y_k - y_{k-1}$, with $j = 1, 2, \ldots, m$ and $k = 1, 2, \ldots, n$, and apply Minkowski's inequality (1.2.7), on p.15 in Sect. 1.2, $n - 1$ times to obtain

$$\left(\sum_{j=1}^{m}\Big|\sum_{i=1}^{n}f(x_j,y_i)\triangle y_i\Big|^p\triangle x_j\right)^{\frac{1}{p}}=\left(\sum_{j=1}^{m}\Big|\sum_{i=1}^{n}f(x_j,y_i)\triangle y_i\,(\triangle x_j)^{\frac{1}{p}}\Big|^p\right)^{\frac{1}{p}}$$

$$\leq\sum_{i=1}^{n}\left(\sum_{j=1}^{m}\Big|f(x_j,y_i)\triangle y_i\,(\triangle x_j)^{\frac{1}{p}}\Big|^p\right)^{\frac{1}{p}}=\sum_{i=1}^{n}\left(\sum_{j=1}^{m}\big|f(x_j,y_i)\big|^p\triangle x_j\right)^{\frac{1}{p}}\triangle y_i.$$

Observe that the first term is a Riemann sum of the integral on the left of (7.6.1) and the last term is a Riemann sum of the integral on the right of (7.6.1). Since the limit of suitable Riemann sums is the Riemann integral, we have established (7.6.1) for continuous functions on bounded rectangles $J_1 \times J_2$. A standard density argument extends continuous functions to functions in L_p and from bounded $J_1 \times J_2$ to unbounded $J_1 \times J_2$.   ∎

For a sequence $\{a_n\}, n = 1, 2, \ldots$ of real numbers, its "limit inferior" is defined by

$$\liminf_{n\to\infty} a_n := \lim_{n\to\infty}\Big\{\inf_{m\geq n} a_m\Big\}.$$

Since $\inf_{m\geq n} a_m = \inf\{a_m : m \geq n\}, n = 1, 2, \ldots$ is a non-decreasing sequence, its limit (which may be $+\infty$) always exists. Clearly,

$$\liminf_{n\to\infty} a_n = \sup_{n\geq 1}\inf_{m\geq n} a_m = \sup\big\{\inf\{a_m : m \geq n\} : n \geq 1\big\}.$$

Similarly, the "limit superior" of $\{a_n\}, n = 1, 2, \ldots$ is defined by

$$\limsup_{n\to\infty} a_n := \lim_{n\to\infty}\Big\{\sup_{m\geq n} a_m\Big\} = \inf\big\{\sup\{a_m : m \geq n\} : n \geq 1\big\}.$$

Clearly $\liminf_{n\to\infty} a_n \leq \limsup_{n\to\infty} a_n$. When $\lim_{n\to\infty} a_n$ exists, then

$$\liminf_{n\to\infty} a_n = \limsup_{n\to\infty} a_n = \lim_{n\to\infty} a_n.$$

Theorem 3 $\boxed{\textbf{Fatou's lemma}}$ *Let $\{f_n(x)\}$ be a sequence of piecewise continuous functions on a finite or infinite interval J with $f_n(x) \geq 0, x \in J, n = 1, 2, \ldots$. Then*

$$\int_J \liminf_{n\to\infty} f_n(x)dx \leq \liminf_{n\to\infty} \int_J f_n(x)dx, \qquad (7.6.2)$$

where the integrals are allowed to be infinite.

Theorem 3 is called Fatou's lemma in Real Analysis. Its proof is beyond the scope of this book. In Real Analysis, the piecewise continuity assumption on $f_n(x), n = 1, 2, \cdots$ is replaced by the general assumption that $f_n(x)$ are measurable functions.

In this case the pointwise limit $f(x) = \lim\limits_{n\to\infty} f_n(x)$, $x \in J$, is also measurable. In this elementary textbook, we assume that $f(x)$ is in $PC(J)$.

Let $\{f_n(x)\}$ be a sequence of piecewise continuous functions on an interval J. If there is $g(x) \in L_1(J)$ such that

$$f_n(x) \le g(x), \quad n = 1, 2, \ldots,$$

for almost all $x \in J$, then applying Fatou's lemma to the sequence $g(x) - f_n(x) \ge 0$, we have

$$\limsup_{n\to\infty} \int_J f_n(x)dx \le \int_J \limsup_{n\to\infty} f_n(x)dx, \qquad (7.6.3)$$

since $\inf\limits_{m\ge n}\{-f_m(x)\} = -\sup\limits_{m\ge n}\{f_m(x)\}$.   ∎

Theorem 4 $\boxed{\textbf{Lebesgue's dominated convergence theorem}}$ *Let J be a finite or infinite interval, and $\{f_n(x)\}$ be a sequence of functions in $L_1(J)$. Suppose that $\lim\limits_{n\to\infty} f_n(x) = f(x)$ for almost all $x \in J$, and that $\{f_n(x)\}$ is dominated by a function $g(x) \in L_1(J)$, namely:*

$$|f_n(x)| \le g(x), \quad n = 1, 2, \ldots, \quad x \in J.$$

Then the order of integration and taking the limit can be interchanged; that is,

$$\lim_{n\to\infty} \int_J f_n(x)dx = \int_J f(x)dx. \qquad (7.6.4)$$

Theorem 4 is also valid for a family of functions $f_h(x) = q(x, h)$ with a continuous parameter h. More precisely, let $a \in \mathbb{R}$ and suppose that $\lim_{h\to a} q(x, h) = f(x)$ for almost all $x \in J$, and that there is $g(x) \in L_1(J)$ and $\delta_0 > 0$ such that

$$\int_J |q(x, h)|dx < \infty, \ |q(x, h)| \le g(x), \ h \in (a - \delta_0, a + \delta_0)$$

for almost all $x \in J$, then

$$\lim_{h\to a} \int_J q(x, h)dx = \int_J f(x)dx.$$

The proof of Theorem 4 is based on Fatou's lemma. Indeed, considering the sequence $|f_n(x) - f(x)|$, $n = 1, 2, \ldots$, we have

$$|f_n(x) - f(x)| \le 2g(x), \quad n = 1, 2, \ldots.$$

Thus by (7.6.3), we have

$$\limsup_{n\to\infty} \int_J |f_n(x) - f(x)|dx \le \int_J \limsup_{n\to\infty} |f_n(x) - f(x)|dx$$

$$= \int_J \lim_{n\to\infty} |f_n(x) - f(x)|dx = 0,$$

where the last equality follows from the assumption $\lim_{n\to\infty} f_n(x) = f(x)$. Thus, $\lim_{n\to\infty} \int_J |f_n(x) - f(x)|dx$ exists and is zero, which implies (7.6.4). ∎

Theorem 5 | **Lebesgue's integration theorem** | *Let* $1 \le p < \infty$. *If* $f \in L_p(\mathbb{R})$, *then*

$$\lim_{t\to 0} \int_{-\infty}^{\infty} |f(x + t) - f(x)|^p dx = 0. \tag{7.6.5}$$

The statement remains valid if $\mathbb{R}$ *is replaced by a finite interval.*

Theorem 5 is concerned with the L_p continuity and is called Lebesgue's integration theorem here. Its proof is provided below.

Proof of Theorem 5 | For an arbitrarily given $\epsilon > 0$, there is a $b > 0$ such that

$$\int_{|x|>b} |f(x)|^p dx < \epsilon, \quad \int_{|x|>b} |f(x + t)|^p dx < \epsilon,$$

where the second inequality holds for all sufficiently small t, say $|t| < 1$. Hence, we may restrict our attention to the interval $[-b - 1, b + 1]$. Since a function $f \in L_p[-b - 1, b + 1]$ can be approximated arbitrarily close by continuous functions, we may assume that f is continuous on $[-b - 1, b + 1]$. Since continuous functions on a closed and bounded interval are uniformly continuous, there exists some δ, $0 < \delta < 1$, such that

$$|f(x) - f(y)| \le \left(\frac{\epsilon}{2b}\right)^{\frac{1}{p}}, \quad \text{for all } x, y \in [-b - 1, b + 1], |x - y| < \delta.$$

Thus for $|t| < \delta$, we have

$$\int_{-b}^{b} |f(x + t) - f(x)|^p dx \le \int_{-b}^{b} \frac{\epsilon}{2b} dx = \epsilon;$$

from which it follows that

$$\int_{-\infty}^{\infty} |f(x + t) - f(x)|^p dx$$

$$= \int_{|x|>b} (|f(x + t)| + |f(x)|)^p dx + \int_{|x|\le b} |f(x + t) - f(x)|^p dx$$

$$\leq \int_{|x|>b} 2^p \big(|f(x+t)|^p + |f(x)|^p\big)dx + \int_{-b}^{b} |f(x+t) - f(x)|^p dx$$

$$\leq 2^p(2\epsilon) + \epsilon = (2^{p+1} + 1)\epsilon,$$

where the inequality $(a+c)^p \leq 2^p(a^p + c^p)$, for any $a, c \geq 0$, has been used in the above derivation. This completes the proof of (7.6.5).   ∎

Chapter 8
Wavelet Transform and Filter Banks
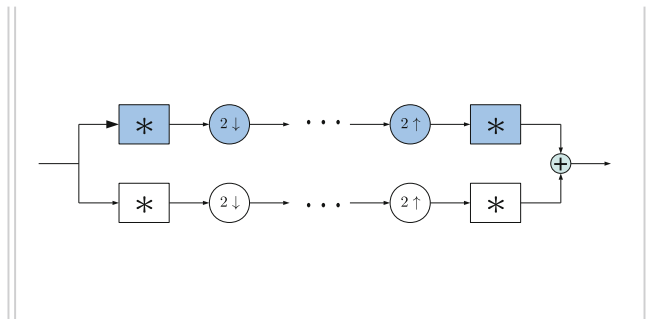
To adopt the Fourier basis functions $e^{i\omega x}$ or e^{inx}, $\cos nx$, and $\sin nx$, where $\omega \in \mathbb{R}$ and $n \in \mathbb{Z}$, for investigating the frequency contents of functions $f(x)$, for $f \in L_1(\mathbb{R})$ or $f \in PC^*[a, b]$, in some desirable neighborhood of any $x = x_0$, called "region of interest (ROI)" in engineering applications, a suitable window function is used to introduce the localized Fourier transform (LFT) and to construct time-frequency basis functions in Sects. 7.4 and 7.5, respectively, of the previous chapter. In this chapter, we consider the approach of replacing the Fourier basis functions by a very general function $\psi \in L_2(\mathbb{R})$ with integral over $(-\infty, \infty)$ equal to zero, such that $\psi(x) \to 0$ for $x \to \pm\infty$. Hence, there is no need to introduce any window function any more. Just as the LFT in Definition 1 on p.351, where the window function $u(t)$ is allowed to slide along the entire "time-domain" $\mathbb{R} = (-\infty, \infty)$, we will also translate $\psi(t)$ along $\mathbb{R}$ by considering $\psi(t - b)$, where $b \in \mathbb{R}$. On the other hand, instead of considering the frequency modulation $u(t-x)e^{-i\omega t}$ of the sliding window $u(t-x)$ in (7.4.1) of Sect. 7.4 as the integral kernel for the LFT operator $\mathbb{F}_u$, a positive parameter $a > 0$, called "scale", is used to introduce the family of functions

$$\psi_{b,a}(t) = \frac{1}{a}\psi\Big(\frac{t-b}{a}\Big),$$

where the normalization by $\frac{1}{a}$ multiplication is used to preserve $L_1(\mathbb{R})$ norm, namely: $||\psi_{b,a}||_1 = ||\psi||_1$ for all $a > 0$ and $b \in \mathbb{R}$.

The functions $\psi_{b,a}(t)$ of the two-parameter family $\{\psi_{b,a}\}$ are called "wavelets" generated by a single wavelet function $\psi \in L_2(\mathbb{R})$. Moreover, analogous to the definition of LFT $\mathbb{F}_u$ on p.351, with integral kernel $u(t - x)e^{-i\omega t}$, the complex conjugate of the wavelet family $\{\psi_{b,a}\}$ is used as the integral kernel to introduce the "wavelet transform" W_ψ, defined by

$$(W_\psi f)(b, a) = \langle f, \psi_{b,a} \rangle = \frac{1}{a} \int_{-\infty}^{\infty} f(t) \overline{\psi\left(\frac{t-b}{a}\right)} dt$$

for functions $f \in L_2(\mathbb{R})$. Hence, the wavelet transform W_ψ may be called "integral wavelet transform", but is commonly called "continuous wavelet transform" (CWT) in the literature.

The significance of the scale parameter $a > 0$ in the wavelet transform $(W_\psi f)(b, a)$ of $f \in L_2(\mathbb{R})$ will be illustrated in Sect. 8.1. In particular, the scale parameter facilitates not only the investigation of the frequency content of $f(x)$, but also the consideration of the "zoom-in" and "zoom-out" functionality. Under the additional assumption that the Fourier transform $\widehat{\psi}(\omega)$ of ψ satisfies the condition

$$|\omega|^{-1} |\widehat{\psi}(\omega)|^2 \in L_1(\mathbb{R}),$$

it will be shown in the same first section that every function $f \in (L_2 \cap L_\infty)(\mathbb{R})$ can be recovered from its wavelet transform $(W_\psi f)(b, a)$ at any $x \in \mathbb{R}$ where f is continuous.

On the other hand, for computational and implementational simplicity, the wavelet transform $(W_\psi f)(b, a)$ is sampled at

$$(b, a) = \left(\frac{k}{2^j}, \frac{1}{2^j}\right), \quad j, k \in \mathbb{Z}.$$

The consideration of the multi-scales 2^{-j}, $j \in \mathbb{Z}$, gives rise to the multi-level structure, called "multiresolution approximation" (MRA), in that a nested sequence of vector subspaces

$$\mathbb{V}_j = \text{closure}_{L_2} \text{span} \left\{ \phi(2^j x - k) : k \in \mathbb{Z} \right\}$$

of $L_2 = L_2(\mathbb{R})$ can be formulated in terms of a single function $\phi(x)$ that satisfies the so-called two-scale relation:

$$\phi(x) = \sum_k p_k \phi(2x - k), \ x \in \mathbb{R},$$

for some sequence $\{p_k\} \in \ell_2$, with the integral of ϕ over $\mathbb{R}$ equal to 1. Here, the subspaces $\mathbb{V}_j$ satisfy the nested and density properties:

$$\{0\} \leftarrow \cdots \subset \mathbb{V}_{-1} \subset \mathbb{V}_0 \subset \mathbb{V}_1 \subset \mathbb{V}_2 \subset \cdots \rightarrow L_2(\mathbb{R}).$$

This will be discussed in Sect. 8.2, with Cardinal B-splines as canonical examples. By considering the complementary subspace $\mathbb{W}_j$ of $\mathbb{V}_{j+1}$ relative to $\mathbb{V}_j$, for each $j \in \mathbb{Z}$, a corresponding wavelet $\psi(x)$, formulated by

$$\psi(x) = \sum_k q_k \phi(2x - k), \ x \in \mathbb{R},$$

can be constructed by finding an appropriate sequence $\{q_k\} \in \ell_2$. Simple examples will be provided in Sect. 8.2 to illustrate the importance of this MRA architecture. Here, together with the wavelet ψ, MRA stands for "multiresolution analysis".

The notion of discrete wavelet transform (DWT) is introduced in Sect. 8.3, by sampling the time and scale parameters of the time-scale coordinate (b, a) for the (integral) wavelet transform at the scales $a = 2^j$ and corresponding time samples $b = k/2^j$, for all integers j and k as discussed above. As a consequence, when a filter pair is used to decompose a bi-infinite signal sequence into low-frequency and high-frequency components, the corresponding bi-orthogonal filter pair can be applied to perfectly reconstruct the signal. This filtering process is called wavelet decomposition and reconstruction, and the same filter pair can be applied to decompose the low-frequency components to as many scale levels as desired, with the same corresponding bi-orthogonal filter pair to reconstruct the decomposed signal components to the previous scale levels. The Haar filter pairs are used to illustrate this multi-level wavelet decomposition/reconstruction algorithm, and the lifting implementation scheme is also demonstrated by using both the Haar filters and the 5/3-tap biorthogonal filters.

The topic of filter banks is studied in Sect. 8.4 to generalize and fine-tune the wavelet decomposition and reconstruction algorithm considered in Sect. 8.3, including the downsampling and upsampling operations and perfect reconstruction. The symbol notation is applied to change the convolution filtering operations to algebraic polynomial multiplications.

8.1 Wavelet Transform

Let us consider a function $\psi \in L_2(\mathbb{R})$ that satisfies $\psi(t) \to 0$ for $t \to \pm\infty$, for which the Cauchy principal value of the integral of ψ on $\mathbb{R}$ exists and vanishes; that is,

$$\text{PV} \int_{-\infty}^{\infty} \psi(t)dt = \lim_{A \to \infty} \int_{-A}^{A} \psi(t)dt = 0. \qquad (8.1.1)$$

Then ψ is called a wavelet; and by introducing two real numbers $b \in \mathbb{R}$ and $a > 0$, we have a two-parameter family of functions

$$\psi_{b,a}(t) = \frac{1}{a}\psi\left(\frac{t - b}{a}\right), \qquad (8.1.2)$$

called wavelets generated by ψ. The word "wavelets" means "small waves" or "ondelettes" in French. In view of (8.1.1), the graph of $\psi(t)$ is oscillating (and hence, "wavy"); and this wave dies down to 0, since $\psi(t) \to 0$ as $|t| \to \infty$. Moreover,

observe that $\psi_{b,a}(t)$ zooms in to a smaller region near $t = b$ as the parameter a tends to 0. Therefore the graphs of $\psi_{b,a}(t)$ are indeed small or large waves, depending on small values or large values of the parameter $a > 0$; and this family of wavelets covers the entire "time-domain" $\mathbb{R} = (-\infty, \infty)$ as b runs over $\mathbb{R}$.

Recall from Sect. 7.4 of the previous chapter that to localize the Fourier transform, a window function $u(t)$ is introduced to define the localized Fourier transform (LFT) $\mathbb{F}_u$ by

$$(\mathbb{F}_u f)(x, \omega) = \langle f, U(\cdot, x, \omega) \rangle = \int_{-\infty}^{\infty} f(t)\overline{U(t, x, \omega)}dt, \qquad (8.1.3)$$

where the complex conjugate of $U(t, x, \omega)$ is defined by

$$\overline{U(t, x, \omega)} = u(t - x)e^{-it\omega} \qquad (8.1.4)$$

(see Definition 1 in Chap. 7 on p.351). In other words, as a function of the variable t, the window function $u(t - x)$ is used to localize the (input) function $f(t)$ near $t = x$ and the modulation by $e^{-it\omega}$ is applied to generate the oscillations in the definition of $\mathbb{F}_u f$ with frequency $= \omega/2\pi\ Hz$ for $\omega \neq 0$.

In contrast to $U(t, x, \omega)$, the wavelets $\psi_{b,a}(t)$ defined in (8.1.2) have both the localization and oscillation features for the analysis of (input) functions $f \in L_2(\mathbb{R})$, when used as the "integration kernel" for the **wavelet transform** of $f(t)$, defined by

$$(W_\psi f)(b, a) = \langle f, \psi_{b,a} \rangle = \frac{1}{a} \int_{-\infty}^{\infty} f(t)\, \overline{\psi\left(\frac{t - b}{a}\right)}\, dt \qquad (8.1.5)$$

for analyzing the oscillation behavior of $f(t)$. The localization feature is easy to understand, since $\psi(t)$ is a window function already. However, it is important to point out that the window size (as defined in terms of the width $2\Delta_\psi$ in Definition 3, (7.4.4) of Chap. 7 on p.352) varies since

$$\Delta_{\psi_{b,a}} = a\Delta_\psi \text{ and } \Delta_{\widehat{\psi}_{b,a}} = \frac{1}{a}\Delta_{\widehat{\psi}}. \qquad (8.1.6)$$

Hence, the wavelet transform $(W_\psi f)(b, a)$ of $f(t)$ zooms in, as the time-window width $2\Delta_{\psi_{b,a}} = 2a\Delta_\psi$ narrows (for smaller values of $a > 0$, with wider frequency-window) and zooms out as the window width widens (for larger values of $a > 0$, with narrower frequency-window for analyzing high-frequency contents).

Next, we must understand the feature of oscillation, or frequency. Since the translation parameter b has nothing to do with oscillation, the frequency must be governed by the scale parameter $a > 0$ as well.

For this purpose, let us consider the single frequency signal $f_\omega(t) = e^{i\omega t}$ with frequency $= \omega/2\pi\ Hz$ (where ω is fixed). Although f_ω is not in $L_2(\mathbb{R})$, the inner product in (8.1.5) is still well-defined [for any wavelet $\psi \in L_2(\mathbb{R})$], since

$$\overline{\langle f_\omega, \psi_{b,a}\rangle} = \int_{-\infty}^{\infty} \psi_{b,a}(t)e^{-it\omega}dt$$

is the Fourier transform of the function $\psi_{b,a} \in L_2(\mathbb{R})$ [see Definition 1 for the definition of Fourier transform for $L_2(\mathbb{R})$ on p.329]. Hence, in view of the properties (ii) and (iii) in Theorem 2 on p.323, we have, from (8.1.5), that

$$\begin{aligned}
\left(W_\psi f_\omega\right)(b, a) = \langle f_\omega, \psi_{b,a}\rangle &= \overline{\left(\mathbb{F}\psi_{b,a}\right)}(\omega) \\
&= \frac{1}{a}\int_{-\infty}^{\infty} \bar{\psi}\left(\frac{t-b}{a}\right) e^{i\omega t}\,dt \\
&= e^{i\omega b}\,\overline{\widehat{\psi}(a\omega)}.
\end{aligned} \tag{8.1.7}$$

Example 1 Apply (8.1.7) to explore the relationship between the notions of "frequency" ω and "scale" a, by considering an appropriate wavelet ψ whose Fourier transform is formulated as the difference of two ideal lowpass filters.

Solution Let us first consider a single-frequency signal

$$g_{\omega_0}(t) = d_0 \cos \omega_0 t$$

with amplitude $= d_0 \neq 0$ and frequency $= \frac{1}{2\pi}\omega_0$ Hz.

Recall from (7.2.14) on p.334 the ideal lowpass filter

$$h_\eta(t) = \frac{\sin \eta t}{\pi t}, \quad \eta > 0,$$

with Fourier transform given by

$$\widehat{h}_\eta(\omega) = \chi_{[-\eta,\eta]}(\omega), \tag{8.1.8}$$

and consider the function $\psi_\epsilon(t)$ defined by

$$\psi_\epsilon(t) = h_{1+\epsilon}(t) - h_{1-\epsilon}(t),$$

with Fourier transform given by

$$\widehat{\psi}_\epsilon(\omega) = \chi_{[-1-\epsilon,-1+\epsilon]}(\omega) + \chi_{(1-\epsilon,1+\epsilon]}(\omega)$$

by applying (8.1.8). Since $0 < \epsilon < 1$, we have $\widehat{\psi}_\epsilon(0) = 0$, or equivalently,

$$\int_{-\infty}^{\infty} \psi_\epsilon(t)dt = \widehat{\psi}_\epsilon(0) = 0,$$

so that $\psi_\epsilon(t)$ is a wavelet [see (8.1.1)]. Now, applying (8.1.7), we have

$$
\begin{aligned}
\left(W_{\psi_\epsilon} g_{\omega_0}\right)(b, a) &= \frac{1}{2} d_0\left(W_{\psi_\epsilon} e^{i\omega_0 t}\right)(b, a) + \frac{1}{2} d_0\left(W_{\psi_\epsilon} e^{-i\omega_0 t}\right)(b, a) \\
&= \frac{1}{2}\, d_0\left(e^{i\omega_0 b} + e^{-i\omega_0 b}\right)\overline{\widehat{\psi_\epsilon}(a\omega_0)} \\
&= d_0\,(\cos\omega_0 b)\left(\chi_{(-1-\epsilon,-1+\epsilon)} + \chi_{(1-\epsilon,1+\epsilon)}\right)(a\omega_0) \\
&= d_0\,(\cos\omega_0 b)\,\chi_{(-1-\epsilon,1+\epsilon)}\,(a\omega_0), \qquad\qquad (8.1.9)
\end{aligned}
$$

since $a\omega_0 > 0$ for positive values of ω_0. An equivalent formulation of (8.1.9) is that

$$
\left(W_{\psi_\epsilon} g_{\omega_0}\right)(b, a) = 0, \ \text{ for } |a\omega_0 - 1| > \epsilon
$$

and

$$
\left(W_{\psi_\epsilon} g_{\omega_0}\right)(b, a) = f_0(b), \ \text{ for } |a\omega_0 - 1| < \epsilon
$$

(where we have ignored the consideration of $|a\omega_0 - 1| = \epsilon$ for convenience). Hence, for small values of $\epsilon > 0$, we see that the relation of the scale a and frequency ω_0 is

$$
\frac{1}{a} \approx \omega_0. \qquad\qquad (8.1.10)
$$

In view of (8.1.10), it is customary to say that the scale a is inversely proportional to the frequency, although this statement is somewhat misleading and only applies to the wavelet $\psi_\epsilon(t)$. ∎

The relationship as illustrated in (8.1.10) between the scale $a > 0$ and frequency can be extended from a single frequency to a range of frequencies, called "frequency band", with bandwidth determined by $\Delta_{\widehat{\psi}}$. (Observe that $\Delta_{\widehat{\psi_\epsilon}} \to 0$ as $\epsilon \to 0$ in Example 1.) To illustrate this concept, let us again consider some wavelet ψ as the difference of two ideal lowpass filters in the following example.

Example 2 Apply (8.1.7) to explore the relationship between the scale $a > 0$ and "frequency bands" $[d^j, d^{j+1})$ for any $d > 1$ and all integers, $j = 0, 1, 2, \ldots$, by considering some wavelet ψ as the difference of two appropriate ideal lowpass filters.

Solution Let us again consider the ideal lowpass filter $h_\eta(t)$ with Fourier transform $\widehat{h}_\eta(\omega) = \chi_{[-\eta,\eta]}(\omega)$ as in (8.1.8) of Example 1, but consider the wavelet

$$
\psi_I(t) = h_d(t) - h_1(t), \qquad\qquad (8.1.11)
$$

where $d > 1$, so that

$$
\widehat{\psi}_I(\omega) = \chi_{[-d,-1)}(\omega) + \chi_{(1,d]}(\omega) \qquad\qquad (8.1.12)
$$

is an "ideal" bandpass filter, with pass-band $= [-d, -1] \cup (1, d]$. Then for a multi-frequency signal

$$g(t) = \sum_{k=0}^{n} c_k \cos kt \tag{8.1.13}$$

with dc term c_0 and ac terms with amplitudes c_k for the frequency components of $k/2\pi\ Hz$, respectively, for $k = 1, \ldots, n$, it follows from the same computation as in Example 1 that

$$(W_{\psi_I} g)(b, a) = \sum_{k=1}^{n} c_k (\cos kb) \chi_{[1,d)}(ak).$$

In particular, for each $j = 0, \ldots, \lfloor \log_d n \rfloor$,

$$(W_{\psi_I} g)\left(b, \frac{1}{d^j}\right) = \sum_{k=1}^{n} c_k (\cos kb) \chi_{[1,d)}\left(\frac{k}{d^j}\right)$$

$$= \sum_{d^j \le k < d^{j+1}} c_k \cos kb. \tag{8.1.14}$$

That is, $(W_{\psi_I} g)(b, d^{-j})$ is precisely the restriction of the given signal $g(t)$ on the "frequency band" $[d^j, d^{j+1})$, where the time variable t of $g(t)$ is replaced by the translation parameter b.

To capture the dc term c_0, we simply use the lowpass filter $h_1(t)$ that generates the wavelet $\psi(t)$, again by taking the inner product, namely,

$$\langle g, h_1 \rangle = \int_{-\infty}^{\infty} g(t)\overline{h_1(t)}dt = \int_{-\infty}^{\infty} g(t)h_1(t)dt$$

$$= \frac{1}{2} \sum_{k=0}^{n} c_k \left(\int_{-\infty}^{\infty} h_1(t)e^{ikt}\,dt + \int_{-\infty}^{\infty} h_1(t)e^{-ikt}\,dt \right)$$

$$= \frac{1}{2} \sum_{k=0}^{n} c_k \left(\chi_{(-1,1)}(-k) + \chi_{(-1,1)}(k) \right)$$

$$= \frac{1}{2}c_0(1 + 1) = c_0. \tag{8.1.15}$$

This result, together with (8.1.14), gives rise to the following decomposition of the given signal $g(t)$ into the frequency bands $[0, 1), [1, d), [d, d^2), \ldots$, namely:

$$g(t) = \langle g, h_1 \rangle + \sum_{j=0}^{\lfloor \log_d n \rfloor} (W_{\psi_I} g)\left(t, \frac{1}{d^j}\right)$$

$$= \langle g, h_1 \rangle + \sum_{j=0}^{\lfloor \log_d n \rfloor} \langle g, \psi_{t,d^{-j}}^I \rangle, \tag{8.1.16}$$

where

$$\psi_{t,d^{-j}}^{I}(x) = d^j \psi_I(d^j(x-t)),$$

and

$$\left\langle g, \psi_{t,d^{-j}}^{I} \right\rangle = (W_{\psi_I} g)(t, d^{-j}),$$

as introduced in (8.1.2) and (8.1.5), respectively, for $b = t$ and $a = d^{-j}$.   ∎

We remark that the decomposition formula (8.1.16) derived in Example 2 should be considered only as an illustration of the concept of wavelet decomposition of signals into frequency sub-bands. For computational efficiency, the translation parameter b is also discretized, namely: $b = k/d^j$, so that for $a = d^{-j}$, we have

$$\psi_{b,a}^{I}(x) = d^j \psi_I\left(d^j\left(x - \frac{k}{d^j}\right)\right) = d^j \psi_I(d^j x - k) \qquad (8.1.17)$$

and

$$(W_{\psi_I} f)(b, a) = (W_{\psi_I} f)\left(\frac{k}{d^j}, \frac{1}{d^j}\right) = d^j \int_{-\infty}^{\infty} f(t) \overline{\psi_I(d^j t - k)} dt. \qquad (8.1.18)$$

In the following, we introduce "Parseval's formula" that will be used to derive the inverse wavelet transform.

Definition 1 *Let $\mathbb{R}_+^2$ denote the upper-half plane $(-\infty, \infty) \times (0, \infty)$. Then for $F(b, a)$ and $G(b, a)$ with $\frac{1}{a} F(b, a)$, $\frac{1}{a} G(b, a) \in L_2(\mathbb{R}_+^2)$, the inner product $\langle F, G \rangle_W$ is defined by*

$$\langle F, G \rangle_W = \int_0^{\infty} \left\{ \int_{-\infty}^{\infty} F(b, a) \overline{G(b, a)} \, db \right\} \frac{da}{a}. \qquad (8.1.19)$$

Furthermore, the vector space with inner product defined by (8.1.19) will be denoted by $L_2(\mathbb{R}_+^2, \frac{da}{a})$ and

$$\|F\|_W = \sqrt{\langle F, F \rangle_W}.$$

Definition 2 *A wavelet $\psi \in L_2(\mathbb{R})$ is said to be admissible, if its Fourier transform $\widehat{\psi}$ satisfies*

$$C_\psi = \int_0^{\infty} \frac{|\widehat{\psi}(\omega)|^2}{\omega} \, d\omega < \infty. \qquad (8.1.20)$$

Theorem 1 *Let $\psi \in L_2(\mathbb{R})$ be an admissible wavelet as introduced in Definition 2. Then for any $f \in L_2(\mathbb{R})$, the wavelet transform $(W_\psi f)(b, a)$ is in $L_2\left(\mathbb{R}_+^2, \frac{db \, da}{a}\right)$.*

Proof Since both f and ψ are in $L_2(\mathbb{R})$ and

$$\widehat{\psi}_{b,a}(\omega) = \widehat{\psi}(a\omega) \, e^{-ib\omega}$$

(by applying (ii) and (iii) of Theorem 2 on p.323), it follows from the definition in (8.1.5) and Parseval's formula (7.2.7) of Theorem 2 on p.331 that

$$(W_\psi f)(b, a) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \widehat{f}(\omega) \,\overline{\widehat{\psi}(a\omega)} \, e^{ib\omega} d\omega. \tag{8.1.21}$$

Hence, by introducing the notation

$$\widehat{F}_a(\omega) = \widehat{f}(\omega) \,\overline{\widehat{\psi}(a\omega)}, \tag{8.1.22}$$

which is an $L_1(\mathbb{R})$ function in view of the Cauchy-Schwarz inequality (see Theorem 3 on p.332), we may apply Theorem 4 on p.27 to conclude that

$$F_a(b) = \left(\mathbb{F}^{-1} \widehat{F}_a\right)(b) = \left(\mathbb{F}^{\#} \widehat{F}_a\right)(b)$$

is well defined, with

$$F_a(b) = (W_\psi f)(b, a)$$

almost everywhere by (8.1.22). Hence, we have

$$\int_0^\infty \left\{ \int_{-\infty}^\infty \left|(W_\psi f)(b, a)\right|^2 db \right\} \frac{da}{a}$$
$$= \int_0^\infty \left\{ \int_{-\infty}^\infty |F_a(b)|^2 db \right\} \frac{da}{a}$$
$$= \int_0^\infty \left\{ \frac{1}{2\pi} \int_{-\infty}^\infty \left|\widehat{F}_a(\omega)\right|^2 d\omega \right\} \frac{da}{a}$$
$$= \frac{1}{2\pi} \int_{-0}^\infty \left\{ \int_{-\infty}^\infty |\widehat{f}(\omega)|^2 \, |\widehat{\psi}(a\omega)|^2 d\omega \right\} \frac{da}{a}$$
$$= \frac{1}{2\pi} \int_{-\infty}^\infty |\widehat{f}(\omega)|^2 \left\{ \int_0^\infty |\widehat{\psi}(a\omega)|^2 \, \frac{da}{a} \right\} d\omega$$
$$= \frac{1}{2\pi} \int_{-\infty}^\infty |\widehat{f}(\omega)|^2 \left\{ \int_0^\infty \frac{|\widehat{\psi}(\xi)|^2}{\xi} d\xi \right\} d\omega$$
$$= C_\psi \frac{1}{2\pi} \|\widehat{f}\|_2^2 = C_\psi \|f\|_2^2 < \infty.$$

In the above derivation, we have applied Parseval's identity to an $L_1(\mathbb{R})$ function $F_a(b)$ which may not be in $L_2(\mathbb{R})$. To be rigorous, we should have truncated $F_a(b)$ with $|b| \leq B$, and take the limit as $B \to \infty$ after applying Parseval's identity to the truncated $F_a(b)$. The detail of this argument is left as an exercise. ∎

Theorem 2 | **Parseval's formula for wavelet transform** | *Let* $\psi \in L_2(\mathbb{R})$ *be an admissible wavelet as defined by* (8.1.20)*. Then*

$$\langle W_\psi f, W_\psi g \rangle_W = C_\psi \langle f, g \rangle, \tag{8.1.23}$$

for all $f, g \in L_2(\mathbb{R})$, where the inner product $\langle\ ,\ \rangle_W$ is defined in (8.1.19) and the constant C_ψ is defined in (8.1.20).

Proof To prove this theorem, we first observe that the left-hand side of (8.1.23) is well defined and finite by applying Theorem 1 and the Cauchy-Schwarz inequality for the inner product $\langle\ ,\ \rangle_W$. Hence, by introducing the notation $\widehat{F}_a(\omega) = \widehat{f}(\omega)\overline{\widehat{\psi}(a\omega)}$ and $\overline{\widehat{G}_a(\omega)} = \overline{\widehat{g}(\omega)}\widehat{\psi}(a\omega)$ as in (8.1.22) and observing that they are $L_2(\mathbb{R})$ functions with inverse Fourier transform given by

$$F_a(b) = (W_\psi f)(b, a),\ \ G_a(b) = (W_\psi g)(b, a)$$

almost everywhere, respectively, we may apply Fubini's theorem to compute

$$\begin{aligned}
\langle W_\psi f, W_\psi g \rangle_W &= \int_0^\infty \left\{ \int_{-\infty}^\infty (W_\psi f)\,(b, a)\, \overline{(W_\psi g)(b, a)} db \right\} \frac{da}{a} \\
&= \int_0^\infty \left\{ \int_{-\infty}^\infty F_a(b)\, \overline{G_a(b)} db \right\} \frac{da}{a} \\
&= \int_0^\infty \left\{ \frac{1}{2\pi} \int_{-\infty}^\infty \widehat{F}_a(\omega)\, \overline{\widehat{G}_a(\omega)} d\omega \right\} \frac{da}{a} \\
&= \int_0^\infty \left\{ \frac{1}{2\pi} \int_{-\infty}^\infty \widehat{f}(\omega)\, \overline{\widehat{g}(\omega)}\, |\widehat{\psi}(a\omega)|^2 d\omega \right\} \frac{da}{a} \\
&= \frac{1}{2\pi} \int_{-\infty}^\infty \widehat{f}(\omega)\, \overline{\widehat{g}(\omega)} \left\{ \int_0^\infty |\widehat{\psi}(a\omega)|^2 \frac{da}{a} \right\} d\omega \\
&= \frac{1}{2\pi} \int_{-\infty}^\infty \widehat{f}(\omega)\, \overline{\widehat{g}(\omega)} \left\{ \int_0^\infty \frac{|\widehat{\psi}(\xi)|^2}{\xi} d\xi \right\} d\omega \\
&= C_\psi \frac{1}{2\pi} \int_{-\infty}^\infty \widehat{f}(\omega)\, \overline{\widehat{g}(\omega)}\, d\omega \\
&= C_\psi \langle f, g \rangle. \qquad\blacksquare
\end{aligned}$$

We are now ready to derive the formula for recovering $f(x)$ from its wavelet transform $(W_\psi f)(b, a)$.

Let $g_\sigma(x)$ be the (normalized) Gaussian function defined in (7.1.13) on p.324 and recall from Theorem 5 on p.327 that for any $f \in L_\infty(\mathbb{R})$,

$$(f * g_\sigma)(x) \to f(x) \tag{8.1.24}$$

as $0 < \sigma \to 0$ at each $x \in \mathbb{R}$ where f is continuous. For each $x \in \mathbb{R}$ where both $f(t)$ and $\psi(\frac{t-b}{a})$ are continuous at $t = x$, in addition to (8.1.24), we also have

$$\left(W_\psi g_\sigma(x - \cdot)\right)(b, a) = \frac{1}{a} \int_{-\infty}^{\infty} g_\sigma(x - t)\overline{\psi\left(\frac{t - b}{a}\right)}dt$$

$$\rightarrow \frac{1}{a} \overline{\psi\left(\frac{x - b}{a}\right)} = \bar{\psi}_{b,a}(x) \qquad (8.1.25)$$

as $0 < \sigma \rightarrow 0$, which yields

$$\langle W_\psi f, W_\psi g_\sigma(x - \cdot)\rangle_W$$
$$= \int_0^\infty \left\{\int_{-\infty}^\infty (W_\psi f)(b, a) \overline{(W_\psi g_\sigma(x - \cdot))}(b, a)db\right\} \frac{da}{a}$$
$$\rightarrow \int_0^\infty \left\{\int_{-\infty}^\infty (W_\psi f)(b, a) \psi_{b,a}(x)db\right\} \frac{da}{a}$$
$$= \int_0^\infty \left\{\int_{-\infty}^\infty (W_\psi f)(b, a) \psi\left(\frac{x - b}{a}\right)db\right\} \frac{da}{a^2}.$$

This, together with (8.1.24), yields the following result.

Theorem 3 **Inverse wavelet transform** *Let $\psi \in L_2(\mathbb{R})$ satisfy (8.1.20). Then for all $f \in (L_2 \cap L_\infty)(\mathbb{R})$,*

$$f(x) = \frac{1}{C_\psi} \int_0^\infty \left\{\int_{-\infty}^\infty (W_\psi f)(b, a)\psi\left(\frac{x - b}{a}\right)db\right\} \frac{da}{a^2}$$
$$= \frac{1}{C_\psi} \int_0^\infty \left\{\int_{-\infty}^\infty \langle f, \psi_{b,a}\rangle \psi_{b,a}(x)db\right\} \frac{da}{a} \qquad (8.1.26)$$

almost everywhere, where C_ψ is given by (8.1.20).

To write out (8.1.26) without using the notation $\psi_{b,a}$, we may set

$$K(x, t, b, a) = \frac{1}{C_\psi} \frac{1}{a^3} \psi\left(\frac{x - b}{a}\right) \overline{\psi\left(\frac{t - b}{a}\right)} \qquad (8.1.27)$$

and apply Fubini's theorem to re-formulate (8.1.26) as

$$f(x) = \int_0^\infty \left\{\int_{-\infty}^\infty \int_{-\infty}^\infty f(t)K(x, t, b, a)dt\ db\right\} da, \qquad (8.1.28)$$

which is a reproduction formula for $L_2 \cap L_\infty(\mathbb{R})$, with wavelet kernel $K(x, t, b, a)$ given in (8.1.27), that may be considered as a "reproducing kernel".

Exercises

Exercise 1 Show that the function $\frac{t}{t^2+1}$ is not integrable on $(-\infty, \infty)$, but has finite Cauchy principal value integral.

Exercise 2 For a wavelet $\psi(t)$ with sufficiently fast decay as $t \to \pm\infty$, show that the size of the time-frequency window of $\psi_{b,a}$ is independent of a and b; that is, $\triangle_{\psi_{b,a}} \triangle_{\widehat{\psi}_{b,a}} = \triangle_\psi \triangle_{\widehat{\psi}}$, for all $a > 0$ and $b \in \mathbb{R}$, where $\widehat{\psi}$ is the Fourier transform of ψ.

Exercise 3 Let f_{ω_0} be a single-frequency signal defined by $f_{\omega_0}(t) = b_0 \sin \omega_0 t$, with magnitude $b_0 \neq 0$. Follow Example 1 to compute the wavelet transform $(W_{\psi_\epsilon} f_{\omega_0})(b, a)$ with the wavelet $\psi_\epsilon(t) = h_{1+\epsilon}(t) - h_{1-\epsilon}(t)$, where $h_\eta(t) = \sin \eta t / \pi t$.

Exercise 4 As a continuation of Exercise 3, compute the wavelet transform $(W_\psi f_{\omega_0})(b, a)$ with the wavelet $\psi(t) = h_d(t) - h_1(t)$. Then compute $(W_\psi f)(b, a)$ for $f(t) = \sum_{k=1}^{n} b_k \sin kt$.

Exercise 5 Fill in the computational details in the proof of Theorem 1.

Exercise 6 Fill in the computational details in the proof of Theorem 2.

Exercise 7 Fill in the details in the proof of Theorem 3.

8.2 Multiresolution Approximation and Analysis

In this section we introduce the notion of multiresolution approximation (MRA) and show how MRA leads to compactly supported wavelets. This section serves as an introduction of the MRA method for the construction of compactly supported wavelets, which is studied in some depth in the next section.

In most applications, particularly in signal and image processing, only band-limited functions are of interest. Hence, although the general theory and method of MRA will be studied in this section and the next two chapters under the framework of $L_2 = L_2(\mathbb{R})$, we will first introduce the concept of MRA and that of the corresponding wavelet bandpass filters, by considering ideal lowpass and ideal bandpass filters, to apply to all band-limited functions.

Let $\widehat{\phi}(\omega) = \chi_{[-\pi,\pi]}(\omega)$ denote the ideal lowpass filter. Then for any band-limited function $f(x)$, there exists a (sufficiently large) positive integer J such that $\widehat{f}(\omega)$ vanishes outside the interval

$$[-2^J \pi, \ 2^J \pi].$$

Hence, the ideal lowpass filter $\widehat{\phi}(2^{-J}\omega)$ becomes an allpass filter of all functions including $f(x)$, with bandwidth $\leq 2^{J+1}\pi$; that is,

$$\widehat{f}(\omega)\widehat{\phi}(2^{-J}\omega) = \widehat{f}(\omega)$$

for all ω, or equivalently

$$(f * \phi_J)(x) = f(x)$$

for all x, where

$$\phi_J(x) = 2^J \phi(2^J x)$$

and the function ϕ, defined by

$$\phi(x) = \frac{\sin \pi x}{\pi x}, \tag{8.2.1}$$

is the inverse Fourier transform of $\widehat{\phi}(\omega) = \chi_{[-\pi,\pi]}(\omega)$. On the other hand, it follows from the Sampling Theorem (see Theorem 6 of Sect. 7.2 in Chap. 7, on p.336) that such functions $f(x)$ (with bandwidth not exceeding $2^{J+1}\pi$) can be recovered from its discrete samples $f\left(\frac{k}{2^J}\right)$, $k \in \mathbb{Z}$, via the formula

$$\begin{aligned} f(x) &= \sum_{k=-\infty}^{\infty} f\left(\frac{k}{2^J}\right) \frac{\sin \pi \left(2^J x - k\right)}{\pi \left(2^J x - k\right)} \\ &= \sum_{k=-\infty}^{\infty} c_k^J \phi \left(2^J x - k\right), \end{aligned} \tag{8.2.2}$$

by applying (8.2.1), where

$$c_k^J = f\left(\frac{k}{2^J}\right).$$

Since the function $f(x) = \phi\left(2^{J-1}x\right)$ has bandwidth $= 2^J \pi$, it follows from (8.2.2) that

$$\phi\left(2^{J-1}x\right) = \sum_{k=-\infty}^{\infty} c_k^J \phi\left(2^J x - k\right), \tag{8.2.3}$$

where, in view of (8.2.1),

$$c_k^J = \phi\left(\frac{2^{J-1}k}{2^J}\right) = \phi\left(\frac{k}{2}\right) = \frac{\sin(\pi k/2)}{k\pi/2},$$

which is independent of J. Hence, we may introduce the sequence $\{p_k\}$, defined by

$$p_k = \frac{\sin(\pi k/2)}{\pi k/2} = \begin{cases} \delta_j, & \text{for } k = 2j, \\ \frac{(-1)^j 2}{(2j+1)\pi}, & \text{for } k = 2j+1, \end{cases} \tag{8.2.4}$$

for all $j \in \mathbb{Z}$, and replace $2^{J-1}x$ by x in (8.2.3) to obtain the identity

$$\phi(x) = \sum_{k=-\infty}^{\infty} p_k \phi(2x - k), \tag{8.2.5}$$

to be called the **two-scale relation** or **refinement equation**, with the governing sequence $\{p_k\}$ in (8.2.4) called the corresponding **two-scale sequence** or **refinement mask** of $\phi(x)$.

For each $j \in \mathbb{Z}$, let

$$\mathbb{V}_j = \text{closure}_{L_2} \text{span} \Big\{ 2^j \phi(2^j x - k) : k \in \mathbb{Z} \Big\}. \qquad (8.2.6)$$

Then it follows from the two-scale relation (8.2.5) that the family $\{\mathbb{V}_j\}$ of vector spaces is a nested sequence

$$\cdots \subset \mathbb{V}_{-1} \subset \mathbb{V}_0 \subset \mathbb{V}_1 \subset \mathbb{V}_2 \subset \cdots . \qquad (8.2.7)$$

The reason for considering the above ideal lowpass filter $\phi(x)$ in (8.2.1) in generating the nested sequence of vector spaces $\mathbb{V}_j$, is that the Fourier transform of the generating functions $2^j \phi \left(2^j x \right)$ is precisely the ideal lowpass filter

$$\widehat{\phi} \left(\frac{\omega}{2^j} \right) = \chi_{[-2^j \pi, 2^j \pi]}(\omega)$$

with pass-band $= \left[-2^j \pi, 2^j \pi \right]$. Hence, for each $j \in \mathbb{Z}$,

$$\widehat{\phi} \left(\frac{\omega}{2^j} \right) - \widehat{\phi} \left(\frac{\omega}{2^{j-1}} \right)$$

is the ideal bandpass filter with pass-band

$$[-2^j \pi, -2^{j-1} \pi] \cup [2^{j-1} \pi, 2^j \pi]. \qquad (8.2.8)$$

Therefore, to separate any function $f_J(x)$ with bandwidth $\leq 2^{J+1} \pi$ into components

$$f_J(x) = f_0(x) + g_0(x) + \cdots + g_{J-1}(x) \qquad (8.2.9)$$

on non-overlapping (ideal) frequency bands; that is,

$$\widehat{f_0}(\omega) = \widehat{f_J}(\omega) \chi_{[-\pi, \pi]}(\omega)$$
$$\widehat{g_0}(\omega) = \widehat{f_J}(\omega) \chi_{[-2\pi, -\pi) \cup (\pi, 2\pi]}(\omega)$$
$$\widehat{g_1}(\omega) = \widehat{f_J}(\omega) \chi_{[-2^2\pi, -2\pi) \cup (2\pi, 2^2\pi]}(\omega)$$
$$\vdots$$
$$\widehat{g_{J-1}}(\omega) = \widehat{f_J}(\omega) \chi_{[-2^J\pi, -2^{J-1}\pi) \cup (2^{J-1}\pi, 2^J\pi]}(\omega),$$

it is required to find an ideal bandpass filter $\psi_I(x)$ with Fourier transform

$$\widehat{\psi_I}(\omega) = \chi_{[-2\pi,-\pi)\cup(\pi,2\pi]}(\omega) = \widehat{\phi}\left(\frac{\omega}{2}\right) - \widehat{\phi}(\omega), \qquad (8.2.10)$$

which yields

$$\widehat{\psi_I}\left(\frac{\omega}{2^j}\right) = \widehat{\phi}\left(\frac{\omega}{2^{j+1}}\right) - \widehat{\phi}\left(\frac{\omega}{2^j}\right) = \chi_{[-2^{j+1}\pi,-2^j\pi)\cup(2^j\pi,2^{j+1}\pi]}(\omega)$$

for $j = 0, \ldots, J - 1$. However, for computational and other reasons, we introduce a phase shift of $\psi_I(x)$ to define the "wavelet"

$$\psi(x) = -2\phi(2x - 1) + \phi\left(x - \frac{1}{2}\right), \qquad (8.2.11)$$

so that

$$\widehat{\psi}(\omega) = -e^{-i\frac{\omega}{2}}\left(\widehat{\phi}\left(\frac{\omega}{2}\right) - \widehat{\phi}(\omega)\right). \qquad (8.2.12)$$

Observe that $|\widehat{\psi}(\omega)| = |\widehat{\psi_I}(\omega)|$ by comparing (8.2.12) with (8.2.10), and hence the separation of $f_J(x)$ into components on ideal frequency bands in (8.2.9) remains valid with only a phase shift of $g_j(x)$ by $-(\pi + \omega/2^j)$, $j = 0, \ldots, J - 1$.

In the definition of $\psi(x)$ in (8.2.11), we remark that $\psi \in \mathbb{V}_1$, with

$$\psi(x) = \sum_{k=-\infty}^{\infty} p_k \phi(2x - (k + 1)) - 2\phi(2x - 1)$$

$$= \sum_{k=-\infty}^{\infty} (p_{k-1} - 2\delta_{k-1})\phi(2x - k), \qquad (8.2.13)$$

by applying (8.2.5). Furthermore, since we have, from (8.2.4), that $p_{2j} = \delta_{2j}$ and

$$p_{1-2j} = \frac{2\sin\frac{(1-2j)\pi}{2}}{(1-2j)\pi} = \frac{-2\sin\frac{(2j-1)\pi}{2}}{-(2j-1)\pi} = p_{2j-1},$$

so that

$$(-1)^k p_{1-k} = \begin{cases} p_{2j-1}, & \text{for } k = 2j, \\ -\delta_{2j}, & \text{for } k = 2j + 1, \end{cases}$$

for all k, it follows from (8.2.13) that ψ as defined in (8.2.11) satisfies the two-scale relation

$$\psi(x) = \sum_{k=-\infty}^{\infty} q_k \phi(2x - k), \qquad (8.2.14)$$

with

$$q_k = (-1)^k p_{1-k}, \ k \in \mathbb{Z}. \qquad (8.2.15)$$

For any function ψ defined on $\mathbb{R}$, we will use the notation

$$\psi_{j,k}(x) = 2^{\frac{j}{2}}\psi(2^j x - k), \ \ j, k \in \mathbb{Z}. \tag{8.2.16}$$

We will also call $\psi \in L_2(\mathbb{R})$ an **orthogonal wavelet** provided that the family $\{\psi_{j,k} : j, k \in \mathbb{Z}\}$ is an orthonormal basis of $L_2(\mathbb{R})$. In this chapter we only consider real-valued wavelets.

The subspaces $\mathbb{V}_j$ of $L_2(\mathbb{R})$ defined in (8.2.6) form the so-called multiresolution approximation (MRA). The construction of compactly supported orthogonal wavelets, to be studied in some depth in Chap. 9, is based on MRA. Before introducing the precise definition of MRA, we recall that for a set $W = \{f_1, f_2, \ldots\}$ of functions in $L_2(\mathbb{R})$, $\overline{\text{span}}\,W$ is the L_2-closure of the linear span by W, namely, $\overline{\text{span}}\,W$ consists of all functions given by

$$\sum_{k=1}^{\infty} c_k f_k(x)$$

which converges in $L_2(\mathbb{R})$, where c_k are constants. Recall from Definition 3 in Sect. 1.4 of Chap. 1, on p.38, that the family $W = \{f_1, f_2, \ldots\}$ is complete in $L_2(\mathbb{R})$ if $\overline{\text{span}}\,W = L_2(\mathbb{R})$.

Definition 1   **Multiresolution approximation (MRA)**   *An* MRA *is a nested sequence* $\{\mathbb{V}_j\}$, $j \in \mathbb{Z}$ *of closed subspaces* $\mathbb{V}_j$ *in* $L_2(\mathbb{R})$ *that satisfy the following conditions:*

(1°) $\mathbb{V}_j \subset \mathbb{V}_{j+1}, j \in \mathbb{Z}$;
(2°) $\bigcap_{j \in \mathbb{Z}} \mathbb{V}_j = \{0\}$;
(3°) $\bigcup_{j \in \mathbb{Z}} \mathbb{V}_j$ *is dense in* $L_2(\mathbb{R})$;
(4°) $f \in \mathbb{V}_j \Leftrightarrow f(2\cdot) \in \mathbb{V}_{j+1}$; *and*
(5°)   *there exists a real-valued function* $\phi \in L_2(\mathbb{R})$ *such that* $\{\phi(\cdot - k) : k \in \mathbb{Z}\}$ *is a Riesz basis of* $\mathbb{V}_0$, *that is,* $\mathbb{V}_0 = \overline{\text{span}}\,\{\phi(\cdot - k) : k \in \mathbb{Z}\}$ *and there exist some constants* $c, C > 0$, *such that*

$$c\sum_{k \in \mathbb{Z}} |c_k|^2 \leq \Big\| \sum_{k \in \mathbb{Z}} c_k \phi(x - k) \Big\|^2 \leq C \sum_{k \in \mathbb{Z}} |c_k|^2, \ \text{for all } \{c_k\} \in \ell^2(\mathbb{Z}). \tag{8.2.17}$$

Observe that by (4°) and (5°), $\mathbb{V}_j = \overline{\text{span}}\{\phi_{j,k} : k \in \mathbb{Z}\}$. Thus we say that ϕ generates the MRA $\{\mathbb{V}_j\}$. A function ϕ in (5°) is called a **scaling function** and $\phi \in L_2(\mathbb{R})$ satisfying (8.2.17) is said to be **stable**. An MRA is called an **orthogonal MRA** if ϕ in (5°) is **orthogonal**, in the sense that the collection of its integer translations is an orthonormal system:

$$\langle \phi, \phi(\cdot - k) \rangle = \delta_k, \ \ k \in \mathbb{Z}. \tag{8.2.18}$$

Conditions (1°), (4°) and (5°) in Definition 1 imply that

$$\phi \in \mathbb{V}_0 \subset \mathbb{V}_1 = \overline{\operatorname{span}}\{\phi(2x - k) : k \in \mathbb{Z}\}.$$

Thus ϕ can be written as (8.2.5) for some $p_k \in \mathbb{R}$. A function ϕ satisfying the refinement equation (8.2.5) is said to be **refinable**, and the sequence $\{p_k\}$ is called the refinement mask of ϕ.

Conversely, suppose that ϕ is stable. Let $\mathbb{V}_0 = \overline{\operatorname{span}}\{\phi(\cdot - k) : k \in \mathbb{Z}\}$, and set $\mathbb{V}_j = \{f : f(2^{-j}\cdot) \in \mathbb{V}_0\} = \overline{\operatorname{span}}\{\phi(2^j \cdot - k) : k \in \mathbb{Z}\}$. Then (4°) holds for $\{\mathbb{V}_j\}$. If, in addition, ϕ satisfies (8.2.5), then (1°) is satisfied. For a compactly supported $\phi \in L_2(\mathbb{R})$, it can be shown that (2°) and (3°) follow from (1°), (4°) and (5°). Consequently, $\{\mathbb{V}_j\}$ is an MRA; and in addition, if ϕ is orthogonal, then $\{\mathbb{V}_j\}$ is an orthogonal MRA.

To summarize, to construct an MRA (or an orthogonal MRA), we need only to construct (compactly supported) stable (or orthogonal) and refinable ϕ.

When ϕ generates an orthogonal MRA, then ψ defined by (8.2.14), where q_k, $k \in \mathbb{Z}$ are given by (8.2.15) or more generally, by (8.2.28) below, is an orthogonal wavelet. A multiresolution approximation, together with the orthogonal wavelet ψ defined by (8.2.14), is called an orthogonal **multiresolution analysis** (abbreviated as MRA also).

Example 1 $\boxed{\textbf{Haar MRA}}$ Let $\varphi_0(x) = \chi_{[0,1)}(x)$. Define for $j \in \mathbb{Z}$,

$$\mathbb{V}_j = \left\{ f(x) = \sum_k c_k \varphi_0 \left(2^j x - k\right) : \{c_k\} \in \ell_2(\mathbb{Z}) \right\}.$$

Then it is easy to see that

$$\mathbb{V}_j = \left\{ f \in L_2(\mathbb{R}) : f(x)\Big|_{\left[\frac{k}{2^j}, \frac{k+1}{2^j}\right)} \text{ is a constant function on } \left[\frac{k}{2^j}, \frac{k+1}{2^j}\right) \right\}.$$

Thus $\mathbb{V}_j \subseteq \mathbb{V}_{j+1}$. The nesting property (4°) of $\{\mathbb{V}_j\}$ can also be verified by the refinement of φ_0:

$$\varphi_0(x) = \varphi_0(2x) + \varphi_0(2x - 1)$$

(see Exercise 1). The density property (3°) of $\{\mathbb{V}_j\}$ follows from the fact that a function $f \in L_2(\mathbb{R})$ can be approximated arbitrarily well by piecewise constant functions. Property (2°) can be verified directly. In addition, φ_0 is orthogonal. Thus $\{\mathbb{V}_j\}$ is an orthogonal MRA ∎

$\boxed{\textbf{Symbol of a sequence}}$ For a sequence $g = \{g_k\}$ of real or complex numbers, its **two-scale symbol**, also called **symbol** for short, is defined by

$$\boxed{G(z) = \frac{1}{2} \sum_k g_k z^k}, \tag{8.2.19}$$

where $z \in \mathbb{C} \backslash \{0\}$, and convergence of the series is not considered. For a finite sequence $g = \{g_k\}$ (or an infinite sequence $g = \{g_k\}$ with only finitely many g_k nonzero), $G(z)$ is given by

$$G(z) = \frac{1}{2} \sum_{k=k_1}^{k_2} g_k z^k$$

for some integers k_1 and k_2. Such a $G(z)$ is called a **Laurent polynomial** of z or z^{-1}. In addition, for $z = e^{-i\omega}$, $G(e^{-i\omega})$ is a sum of a finite number of terms of $g_k e^{-ik\omega}$, and is therefore a **trigonometric polynomial**. When the finite sequence $\{g_k\}$ is used as a convolution digital filter, then $2G(z)$ is called the z-transform of $\{g_k\}$. Since the z-transform of a digital filter is the transform function of the filtering process, we sometimes call $G(z)$ or $G(e^{-i\omega})$ a **finite impulse response** (**FIR**) filter. ∎

Remark 1 To distinguish the two-scale symbol notation for the refinement and wavelet sequences $p = \{p_k\}$, $\widetilde{p} = \{\widetilde{p}_k\}$ and $q = \{q_k\}$, $\widetilde{q} = \{\widetilde{q}_k\}$ from the z-transform notation of a sequence, we do not bold-face $p, \widetilde{p}, q, \widetilde{q}$ throughout this book. ∎

Taking the Fourier transform of both sides of (8.2.5) and using the fact that the Fourier transform of $\phi(2x - k)$ is $\frac{1}{2}\widehat{\phi}(\omega)e^{-i\omega/2}$, we may deduce that (8.2.5) can be written as

$$\widehat{\phi}(\omega) = P\left(e^{-i\frac{\omega}{2}}\right) \widehat{\phi}\left(\frac{\omega}{2}\right), \tag{8.2.20}$$

where $P(z)$ is the two-scale symbol of $\{p_k\}$:

$$P(z) = \frac{1}{2} \sum_k p_k z^k.$$

Hence, applying (8.2.20) repeatedly, we have

$$\widehat{\phi}(\omega) = P\left(e^{-i\omega/2}\right) P\left(e^{-i\omega/2^2}\right) \widehat{\phi}\left(\frac{\omega}{2^2}\right)$$

$$= \cdots$$

$$= P\left(e^{-i\omega/2}\right) P\left(e^{-i\omega/2^2}\right) \cdots P\left(e^{-i\omega/2^n}\right) \widehat{\phi}\left(\frac{\omega}{2^n}\right)$$

$$= \prod_{j=1}^{n} P\left(e^{-i\omega/2^j}\right) \widehat{\phi}\left(\frac{\omega}{2^n}\right).$$

Suppose that the Laurent polynomial $P(z)$ satisfies $P(1) = 1$, then $\prod_{j=1}^{n} P\left(e^{-i\omega/2^j}\right)$ converges pointwise on $\mathbb{R}$. Thus ϕ, defined by

$$\widehat{\phi}(\omega) = \prod_{j=1}^{\infty} P\left(e^{-i\omega/2^j}\right), \qquad (8.2.21)$$

is a solution of (8.2.20) with $\widehat{\phi}(0) = 1$. In addition, $\text{supp}\phi \subseteq [N_1, N_2]$, provided that $\text{supp}p \subseteq [N_1, N_2]$; that is, $p_k = 0$ for $k < N_1$ or $k > N_2$. This ϕ is called the **normalized solution** of (8.2.5). A detailed discussion of these properties of ϕ will be provided in Chaps. 9 and 10. Furthermore, in Chap. 10, we will derive the condition on $p = \{p_k\}$ so that ϕ is in $L_2(\mathbb{R})$ and that ϕ is orthogonal. From now on, we will always assume that a refinement mask $p = \{p_k\}$ satisfies $P(1) = 1$, i.e.,

$$\boxed{\sum_k p_k = 2};$$

and that the function ϕ associated with $p = \{p_k\}$ means the normalized solution of (8.2.5) given by (8.2.21).

Example 2 $\boxed{\textbf{Refinement of Cardinal B-splines}}$ Let $\varphi_0(x) = \chi_{[0,1)}(x)$ as introduced in Example 1. Then $\phi = \varphi_0 * \varphi_0$ is the piecewise linear B-spline, also called the **hat function**, given by

$$\phi(x) = \min\{x, 2 - x\}\chi_{[0,2)}(x) = \begin{cases} x, & \text{for } 0 \le x < 1, \\ 2 - x, & \text{for } 1 \le x < 2, \\ 0, & \text{elsewhere.} \end{cases} \qquad (8.2.22)$$

From

$$\widehat{\varphi_0}(\omega) = \frac{1 - e^{-i\omega}}{i\omega}$$

(see Exercise 1 in Sect. 7.1 on p.327 and (v) in Theorem 2 of Sect. 7.1 on p.323), we have

$$\begin{aligned}
\widehat{\phi}(\omega) = \widehat{\varphi_0}(\omega)^2 &= \left(\frac{1 - e^{-i\omega}}{i\omega}\right)^2 \\
&= \left(\frac{1 + e^{-i\omega/2}}{2}\right)^2 \left(\frac{1 - e^{-i\omega/2}}{i\omega/2}\right)^2 \\
&= \left(\frac{1 + e^{-i\omega/2}}{2}\right)^2 \widehat{\phi}\left(\frac{\omega}{2}\right).
\end{aligned}$$

Thus, ϕ is refinable with the symbol of the refinement mask $P(z)$ given by

$$P(z) = \left(\frac{1 + z}{2}\right)^2,$$
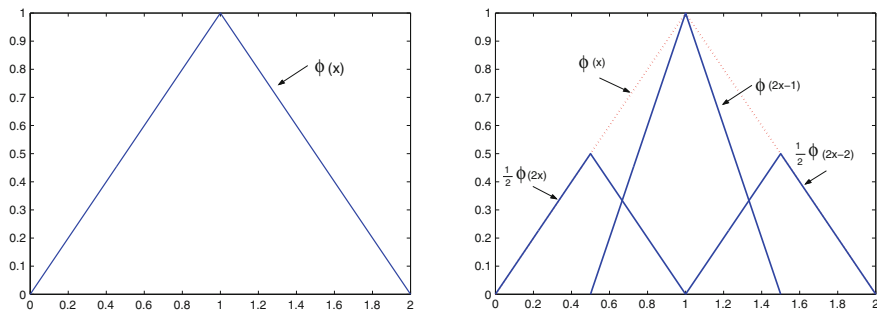
Fig. 8.1. Hat function ϕ (on *left*) and its refinement (on *right*)

or the refinement mask is given by

$$p_0 = \frac{1}{2}, \; p_1 = 1, \; p_2 = \frac{1}{2}, \;\; p_k = 0, k \neq 0, 1, 2;$$

that is,

$$\phi(x) = \frac{1}{2}\phi(2x) + \phi(2x - 1) + \frac{1}{2}\phi(2x - 2)$$

(see Fig. 8.1 for the refinement of ϕ).

More generally, let ϕ be the Cardinal B-spline of order m defined by

$$\phi(x) = \underbrace{(\varphi_0 * \varphi_0 * \cdots * \varphi_0)}_{m \text{ copies of } \varphi_0}(x) \tag{8.2.23}$$

(see Exercise 12 in Sect. 7.1 of Chap. 7 on p.328, where $\varphi_0(x) = h_1(x)$). Then ϕ is refinable with the symbol of the refinement mask given by (see Exercise 3)

$$P(z) = \left(\frac{1 + z}{2}\right)^m. \tag{8.2.24}$$

∎

To construct an orthogonal wavelet, we start with a lowpass FIR filter $p = \{p_k\}$ such that the corresponding scaling function ϕ is orthogonal; that is, the integer shifts $\phi(x - k)$, $k \in \mathbb{Z}$ of ϕ are orthogonal to each other. This implies that the corresponding symbol $P(z)$ must satisfy

$$|P(z)|^2 + |P(-z)|^2 = 1, \; z \neq 0. \tag{8.2.25}$$

Such a filter p (or $P(z)$) is called a **quadrature mirror filter (QMF)**. By applying $P(z)$, we construct the corresponding highpass filter $q = \{q_k\}$ under the construction

criteria

$$|Q(z)|^2 + |Q(-z)|^2 = 1, \ z \neq 0, \tag{8.2.26}$$

$$P(z)Q\left(\frac{1}{z}\right) + P(-z)Q\left(-\frac{1}{z}\right) = 0, \ z \neq 0. \tag{8.2.27}$$

Then ψ, defined by (8.2.14), is a compactly supported orthogonal wavelet. The detailed discussion is provided in the next chapter. Here, we only mention that the highpass filter can be given by the corresponding QMF as shown below.

 Choice of highpass filter q for a QMF lowpass filter p For a Laurent polynomial $P(z) = \frac{1}{2}\sum_k p_k z^k$ with real-valued coefficients p_k, define the Laurent polynomial $Q(z) = \frac{1}{2}\sum_k q_k z^k$ by

$$Q(z) = -z^{2L-1}P\left(-\frac{1}{z}\right),$$

or equivalently by

$$q_k = (-1)^k p_{2L-1-k}, \ k \in \mathbb{Z}, \tag{8.2.28}$$

where L is any desirable integer. It can be verified that if $P(z)$ is a QMF, then $Q(z)$ satisfies (8.2.26) and (8.2.27) (see Exercise 10). If we choose $L = 1$, then (8.2.28) is reduced to (8.2.15). ∎

Example 3 **Haar wavelet** As a continuation of Example 1, let $\varphi_0(x) = \chi_{[0,1)}(x)$ with its refinement mask $p_0 = p_1 = 1$ and $p_k = 0$ for $k \neq 0, 1$. Define q_k by $q_k = (-1)^k p_{1-k}$, namely:

$$q_0 = 1, \ q_1 = -1; \ q_k = 0, k \neq 0, 1.$$

Then the corresponding wavelet ψ given by

$$\psi(x) = \varphi_0(2x) - \varphi_0(2x - 1) = \begin{cases} 1, & \text{for } 0 \leq x < \frac{1}{2}, \\ -1, & \text{for } \frac{1}{2} \leq x < 1, \\ 0, & \text{elsewhere} \end{cases} \tag{8.2.29}$$

is an orthogonal wavelet to be called the Haar wavelet. ∎

Example 4 **D_4 orthogonal wavelet** Let $\{p_k\}$ be the filter with the only nonzero terms p_k given by

$$p_0 = \frac{1 - \sqrt{3}}{4}, \ p_1 = \frac{3 - \sqrt{3}}{4}, \ p_2 = \frac{3 + \sqrt{3}}{4}, \ p_3 = \frac{1 + \sqrt{3}}{4}.$$

Then the symbol $P(z)$ of $\{p_k\}$ is given by

$$P(z) = \frac{1-\sqrt{3}}{8} + \frac{3-\sqrt{3}}{8}z + \frac{3+\sqrt{3}}{8}z^2 + \frac{1+\sqrt{3}}{8}z^3, \tag{8.2.30}$$

and $P(z)$ is a QMF (see Exercise 11). Define $q_k = (-1)^k p_{3-k}$. Then the only nonzero q_k of the sequence $\{q_k\}$ are given by

$$q_0 = \frac{1+\sqrt{3}}{4}, \ q_1 = -\frac{3+\sqrt{3}}{4}, \ q_2 = \frac{3-\sqrt{3}}{4}, \ q_3 = -\frac{1-\sqrt{3}}{4}.$$

Thus, the highpass filter $Q(z)$ for the orthogonal wavelet is given by

$$Q(z) = \frac{1+\sqrt{3}}{8} - \frac{3+\sqrt{3}}{8}z + \frac{3-\sqrt{3}}{8}z^2 - \frac{1-\sqrt{3}}{8}z^3. \tag{8.2.31}$$

Then $Q(z)$ satisfies (8.2.26) and $P(z)$, $Q(z)$ satisfy (8.2.27) (see Exercise 11).

Observe that $P(z)$ has four nonzero terms. Thus this filter is called a 4-tap orthogonal filter, or 4-tap Daubechies filter (D_4 filter for short). See Fig. 8.2 for the corresponding scaling function ϕ and orthogonal wavelet ψ. A detailed discussion on the construction of compactly supported orthogonal wavelets is provided in Sect. 9.3 of the next chapter. ∎

A compactly supported orthogonal wavelet has the limitation that it cannot be symmetric unless it is the Haar wavelet. To attain the symmetry property, the orthogonality property is relaxed to biorthogonality. Two compactly supported functions ψ and $\widetilde{\psi}$ are called a pair of **biorthogonal wavelets** if two families $\{\psi_{j,k}\}_{j,k\in\mathbb{Z}}$ and $\{\widetilde{\psi}_{j,k}\}_{j,k\in\mathbb{Z}}$ are biorthogonal to each other; that is,

$$\langle \psi_{j,k}, \widetilde{\psi}_{j',k'} \rangle = \delta_{j-j'}\delta_{k-k'}, \ j, k, j', k' \in \mathbb{Z}, \tag{8.2.32}$$

and each of the families is a Riesz basis of $L_2(\mathbb{R})$. In this case, one of the pair $\{\psi, \widetilde{\psi}\}$ is called a **dual** of the other.
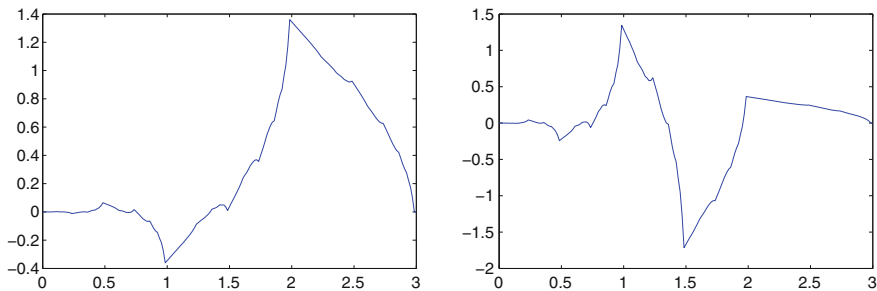
Fig. 8.2. D_4 scaling function ϕ (on *left*) and wavelet ψ (on *right*)

The construction of compactly supported biorthogonal wavelets starts with two FIR lowpass filters $p = \{p_k\}$ and $\widetilde{p} = \{\widetilde{p}_k\}$, such that the corresponding scaling functions ϕ and $\widetilde{\phi}$ are biorthogonal to each other, in the sense that

$$\langle \phi(\cdot - j), \widetilde{\phi}(\cdot - k) \rangle = \delta_{j-k}, \quad j, k \in \mathbb{Z}.$$

This requires that the symbols $P(z)$ and $\widetilde{P}(z)$ of p and $\widetilde{p}$ satisfy

$$\widetilde{P}(z) P\left(\frac{1}{z}\right) + \widetilde{P}(-z) P\left(-\frac{1}{z}\right) = 1. \tag{8.2.33}$$

Then we choose two FIR filters $Q(z)$ and $\widetilde{Q}(z)$ that meet the construction criteria:

$$\widetilde{Q}(z) Q\left(\frac{1}{z}\right) + \widetilde{Q}(-z) Q\left(-\frac{1}{z}\right) = 1, \tag{8.2.34}$$

$$\widetilde{Q}(z) P\left(\frac{1}{z}\right) + \widetilde{Q}(-z) P\left(-\frac{1}{z}\right) = 0, \tag{8.2.35}$$

$$\widetilde{P}(z) Q\left(\frac{1}{z}\right) + \widetilde{P}(-z) Q\left(-\frac{1}{z}\right) = 0. \tag{8.2.36}$$

Consequently, we may define the wavelets ψ and $\widetilde{\psi}$ by their Fourier transforms

$$\widehat{\psi}(\omega) = Q\left(e^{-i\frac{\omega}{2}}\right) \widehat{\phi}\left(\frac{\omega}{2}\right), \quad \widehat{\widetilde{\psi}}(\omega) = \widetilde{Q}\left(e^{-i\frac{\omega}{2}}\right) \widehat{\widetilde{\phi}}\left(\frac{\omega}{2}\right),$$

or equivalently, directly by

$$\psi(x) = \sum_k q_k \phi(2x - k), \quad \widetilde{\psi}(x) = \sum_k \widetilde{q}_k \widetilde{\phi}(2x - k).$$

Then ϕ, $\widetilde{\phi}$, ψ and $\widetilde{\psi}$ satisfy

$$\langle \psi, \widetilde{\psi}(\cdot - k) \rangle = \delta_k, \quad k \in \mathbb{Z}; \tag{8.2.37}$$

$$\langle \phi, \widetilde{\psi}(\cdot - k) \rangle = 0, \quad \langle \widetilde{\phi}, \psi(\cdot - k) \rangle = 0, \quad k \in \mathbb{Z}; \tag{8.2.38}$$

and $\{\mathbb{V}_j\}$ and $\{\widetilde{\mathbb{V}}_j\}$, generated by ϕ and $\widetilde{\phi}$, respectively, as in

$$\mathbb{V}_j = \overline{\text{span}}\{\phi(2^j \cdot -k) : k \in \mathbb{Z}\}, \quad \widetilde{\mathbb{V}}_j = \overline{\text{span}}\{\widetilde{\phi}(2^j \cdot -k) : k \in \mathbb{Z}\},$$

form two MRAs which are biorthogonal to each other, in the sense that

$$\langle \phi_{j,k}, \widetilde{\phi}_{j,k'} \rangle = \delta_{k-k'}, \quad j, k, k' \in \mathbb{Z}.$$

Furthermore, with $\{\mathbb{W}_j\}$ and $\{\widetilde{\mathbb{W}}_j\}$ defined by

$$\mathbb{W}_j = \overline{\mathrm{span}}\{\psi(2^j \cdot -k) :\ k \in \mathbb{Z}\},\ \widetilde{\mathbb{W}}_j = \overline{\mathrm{span}}\{\widetilde{\psi}(2^j \cdot -k) :\ k \in \mathbb{Z}\},$$

we have

$$\mathbb{V}_{j+1} = \mathbb{W}_j + \mathbb{V}_j,\quad \widetilde{\mathbb{V}}_{j+1} = \widetilde{\mathbb{W}}_j + \widetilde{\mathbb{V}}_j. \tag{8.2.39}$$

This, together with $\mathbb{V}_j \perp \widetilde{\mathbb{W}}_j$ and $\widetilde{\mathbb{V}}_j \perp \mathbb{W}_j$ (as follows from (8.2.38)), implies

$$\mathbb{W}_j \perp \widetilde{\mathbb{W}}_{j'},\ \text{ for any } j \neq j'.$$

Hence, ψ and $\widetilde{\psi}$ constructed in this way satisfy (8.2.32), and they form a pair of compactly supported biorthogonal wavelets. A detailed discussion on the construction of biorthogonal wavelets is provided in Sect. 9.4 of the next chapter.

If p and $\widetilde{p}$ satisfy (8.2.33), then we say that they are biorthogonal to each other; and if p, q and $\widetilde{p}$, $\widetilde{q}$ satisfy (8.2.33)–(8.2.36), then we say that they are biorthogonal or constitute a biorthogonal filter bank. The filters p and $\widetilde{p}$ (or $P(z)$ and $\widetilde{P}(z)$) are called lowpass filters; and q and $\widetilde{q}$ (or $Q(z)$ and $\widetilde{Q}(z)$) are called highpass filters.

Choices of highpass filters q, $\widetilde{q}$ for biorthogonal filters p, $\widetilde{p}$ For a given pair of biorthogonal FIR lowpass filters p and $\widetilde{p}$, we may choose the FIR highpass filters q and $\widetilde{q}$ as follows:

$$Q(z) = -z^{2s-1} \widetilde{P}\left(-\frac{1}{z}\right),\quad \widetilde{Q}(z) = -z^{2s-1} P\left(-\frac{1}{z}\right), \tag{8.2.40}$$

where s is any desirable integer. It can be verified that if p and $\widetilde{p}$ are biorthogonal; that is, $P(z)$ and $\widetilde{P}(z)$ satisfy (8.2.33), then $Q(z)$ and $\widetilde{Q}(z)$ satisfy (8.2.34)–(8.2.36) (see Exercise 12). That is, p, q, $\widetilde{p}$, $\widetilde{q}$ constitute a biorthogonal filter bank. ∎

Example 5 **5/3-tap biorthogonal filter bank** Let $\{p_k\}$, $\{\widetilde{p}_k\}$ be the filters given by

$$P(z) = \frac{1}{4} + \frac{1}{2}z + \frac{1}{4}z^2,$$

$$\widetilde{P}(z) = -\frac{1}{8}z^{-1} + \frac{1}{4} + \frac{3}{4}z + \frac{1}{4}z^2 - \frac{1}{8}z^3. \tag{8.2.41}$$

Also, let $\{q_k\}$, $\{\widetilde{q}_k\}$ be the filters defined by (8.2.40) with $s = 1$. Then

$$Q(z) = -\frac{1}{8}z^{-2} - \frac{1}{4}z^{-1} + \frac{3}{4} - \frac{1}{4}z - \frac{1}{8}z^2,$$

$$\widetilde{Q}(z) = -\frac{1}{4}z^{-1} + \frac{1}{2} - \frac{1}{4}z. \tag{8.2.42}$$
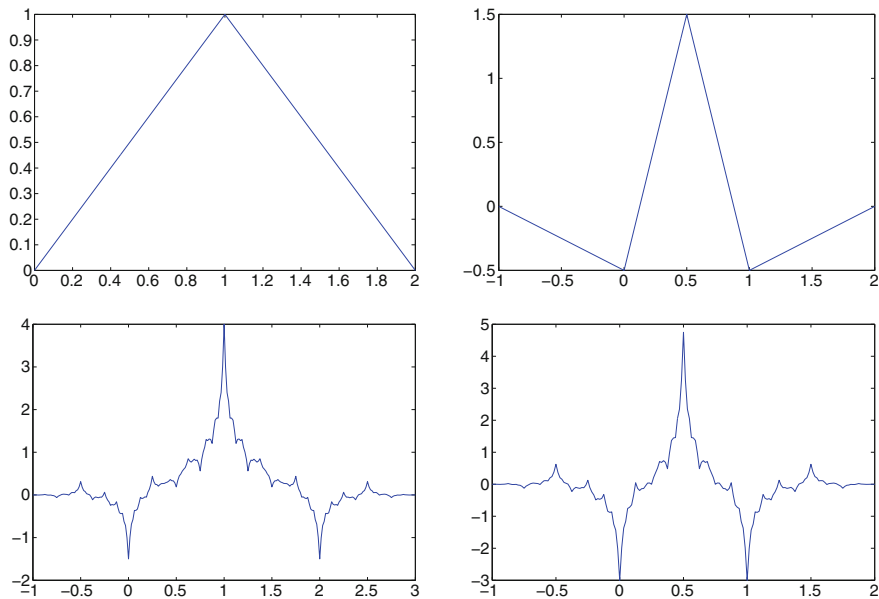
Fig. 8.3. *Top*: biorthogonal 5/3 scaling function ϕ (on *left*) and wavelet ψ (on *right*). *Bottom*: scaling function $\widetilde{\phi}$ (on *left*) and wavelet $\widetilde{\psi}$ (on *right*)

Then $P(z), Q(z), \widetilde{P}(z), \widetilde{Q}(z)$ satisfy (8.2.34)–(8.2.36) (see Exercise 13); that is, $p, q, \widetilde{p}, \widetilde{q}$ constitute a biorthogonal filter bank.

Observe that the scaling function associated with $P(z)$ in (8.2.41) is the linear B-spline (i.e. the hat function). These two lowpass filters $\widetilde{p}$ and p have 5 and 3 nonzero terms, respectively, and they are commonly called the 5/3-tap biorthogonal filters. The 5/3-tap biorthogonal filter pairs have been adopted by the JPEG2000 image compression standard for lossless digital image compression. The graphs of the corresponding scaling functions and biorthogonal wavelets are displayed in Fig. 8.3. ■

Exercises

Exercise 1 Let $\{\mathbb{V}_j\}$ be the sequence of closed spaces defined in Example 1. Show that $\mathbb{V}_j \subset \mathbb{V}_{j+1}$ for all $j \in \mathbb{Z}$.

Exercise 2 Let $\{\mathbb{V}_j\}$ be the sequence of closed spaces defined in Example 1. Show directly that $\cap_{j \in \mathbb{Z}} \mathbb{V}_j = \{0\}$.

Exercise 3 Let ϕ be the Cardinal B-spline of order m defined by (8.2.23). Show that ϕ is refinable with refinement mask given by (8.2.24) (see Exercise 12 in Sect. 7.1 of Chap. 7 on p.328, where $\varphi_0(x) = h_1(x)$).

Exercise 4 Decide whether $\phi(x) = \chi_{[0,1.5)}(x)$ is refinable or not.

Exercise 5 Let $\phi(x)$ be the hat function defined by (8.2.22), and $\varphi(x) = \phi(x) + \phi(x - 1)$. Show that $\varphi(x)$ is refinable with

$$\varphi(x) = \frac{1}{2}\varphi(2x) + \frac{1}{2}\varphi(2x - 1) + \frac{1}{2}\varphi(2x - 2) + \frac{1}{2}\varphi(2x - 3).$$

Exercise 6 Let ϕ be the hat function defined by (8.2.22). Show directly that

$$\phi(x) = \frac{1}{2}\phi(2x) + \phi(2x - 1) + \frac{1}{2}\phi(2x - 2), \ x \in \mathbb{R}.$$

Hint: By the symmetry of ϕ, it is sufficient to consider $x \in (-\infty, 1]$; and then, consider the separate cases $x \in (-\infty, 0), [0, 0.5), [0.5, 1]$.

Exercise 7 Let $\phi = \varphi_0 * \varphi_0 * \varphi_0$ be the quadratic B-spline defined by

$$\phi(x) = \begin{cases} \frac{1}{2}x^2, & \text{if } 0 \le x < 1, \\ -(x-1)^2 + (x-1) + \frac{1}{2}, & \text{if } 1 \le x < 2, \\ \frac{1}{2}\big(1 - (x-2)\big)^2, & \text{if } 2 \le x < 3, \\ 0, & \text{elsewhere.} \end{cases}$$

Show directly that

$$\phi(x) = \frac{1}{4}\phi(2x) + \frac{3}{4}\phi(2x - 1) + \frac{3}{4}\phi(2x - 2) + \frac{1}{4}\phi(2x - 3), \ x \in \mathbb{R}.$$

Hint: By the symmetry of ϕ, it is enough to consider $x \in (-\infty, 1.5]$; and then, consider the separate cases $x \in (-\infty, 0), [0, 0.5), [0.5, 1), [1, 1.5]$.

Exercise 8 Decide whether $\phi(x) = \chi_{[0,1)}(x) + \frac{1}{2}\chi_{[1,2)}(x)$ is refinable or not.

Exercise 9 Let $\varphi(x) = \phi(x) + \frac{1}{2}\phi(x - 1)$, where $\phi(x)$ is the hat function defined by (8.2.22). Decide whether $\varphi(x)$ is refinable or not.

Exercise 10 Suppose that the FIR lowpass filter $P(z)$ is a QMF. Show that $Q(z)$ defined by (8.2.28) satisfies (8.2.26) and (8.2.27).

Exercise 11 Let $P(z)$ and $Q(z)$ be the D_4 orthogonal filters given by (8.2.30) and (8.2.31). Verify directly that (i) $P(z)$ satisfies (8.2.25); (ii) $Q(z)$ satisfies (8.2.26); and (iii) $P(z), Q(z)$ satisfy (8.2.27).

Exercise 12 Suppose that the FIR filters $P(z)$ and $\widetilde{P}(z)$ satisfy (8.2.33). Show that $Q(z)$ and $\widetilde{Q}(z)$ defined by (8.2.40) satisfy (8.2.34)–(8.2.36).

Exercise 13 Let $P(z), \widetilde{P}(z), Q(z), \widetilde{Q}(z)$ be the filters defined by (8.2.41) and (8.2.42). Verify directly that they satisfy (8.2.33)–(8.2.36).

Exercise 14 Show that $\mathbb{W}_j \perp \widetilde{\mathbb{W}}_{j'}$ for $j \ne j'$, and apply this fact to validate (8.2.32).

8.3 Discrete Wavelet Transform

Suppose that $p, q, \widetilde{p}, \widetilde{q}$ constitute a biorthogonal FIR filter bank, in that their symbols satisfy (8.2.33)–(8.2.36) in the previous section. Let $\phi, \widetilde{\phi}$ be the scaling functions associated with the lowpass filters p and $\widetilde{p}$, and let $\psi, \widetilde{\psi}$ be the corresponding biorthogonal wavelets defined by $q, \widetilde{q}$. Also, let $\{\mathbb{V}_j\}$ and $\{\widetilde{\mathbb{V}}_j\}$ be the MRAs generated by ϕ and $\widetilde{\phi}$, respectively. Assume that a real-valued function $f(x)$ is in $\mathbb{V}_j$ for some j, say $j = 0$. Then $f(x)$ can be expanded as

$$f(x) = \sum_k x_k \phi(x - k), \tag{8.3.1}$$

for some real numbers x_k. On the other hand, by (8.2.39) with $j = 0$ in the previous section, on p.402, $f(x)$ can also be expanded as

$$f(x) = \sum_k c_k \phi\left(\frac{x}{2} - k\right) + \sum_k d_k \psi\left(\frac{x}{2} - k\right), \tag{8.3.2}$$

for some real numbers c_k, d_k.

Multiplying both sides of (8.3.2) by $\frac{1}{2}\widetilde{\phi}\left(\frac{x}{2} - n\right)$ and integrating, we may deduce from the biorthogonality of $\widetilde{\phi}$ with respect to ϕ, ψ (see (8.2.38) in the previous section on p.401), that for any $n \in \mathbb{Z}$,

$$\int_{-\infty}^{\infty} f(x) \frac{1}{2}\widetilde{\phi}\left(\frac{x}{2} - n\right) dx = \sum_k c_k \, \delta_{k-n} + \sum_k d_k \, 0 = c_n.$$

In the above equality, we have used the fact that $\widetilde{\phi}$ is real-valued. This property of $\widetilde{\phi}$ will also be used below. Therefore, from (8.3.1), we obtain

$$
\begin{aligned}
c_n &= \frac{1}{2} \int_{-\infty}^{\infty} f(x) \widetilde{\phi}\left(\frac{x}{2} - n\right) dx \\
&= \frac{1}{2} \int_{-\infty}^{\infty} \sum_k x_k \phi(x - k) \, \widetilde{\phi}\left(\frac{x}{2} - n\right) dx \\
&= \frac{1}{2} \sum_k x_k \int_{-\infty}^{\infty} \phi(x - k) \, \widetilde{\phi}\left(\frac{x}{2} - n\right) dx.
\end{aligned}
$$

By the refinement of $\widetilde{\phi}$, we have that

$$\widetilde{\phi}\left(\frac{x}{2} - n\right) = \sum_\ell \widetilde{p}_\ell \widetilde{\phi}(x - 2n - \ell).$$

Thus,

$$c_n = \frac{1}{2} \sum_k x_k \sum_\ell \widetilde{p}_\ell \int_{-\infty}^{\infty} \phi(x-k)\, \widetilde{\phi}(x-2n-\ell) dx$$

$$= \frac{1}{2} \sum_k x_k \sum_\ell \widetilde{p}_\ell \delta_{k-(2n+\ell)} = \frac{1}{2} \sum_k x_k \widetilde{p}_{k-2n}.$$

Similarly, it can be shown that

$$d_n = \frac{1}{2} \sum_k x_k \widetilde{q}_{k-2n}. \tag{8.3.3}$$

Hence, the sequence $\{x_k\}$ can be decomposed into two sequences $\{c_k\}$ and $\{d_k\}$. On the other hand, from (8.3.2), we have that

$$x_n = \langle f, \widetilde{\phi}(\cdot - n) \rangle$$

$$= \sum_k c_k \langle \phi\left(\frac{\cdot}{2} - k\right), \widetilde{\phi}(\cdot - n) \rangle + \sum_k d_k \langle \psi\left(\frac{\cdot}{2} - k\right), \widetilde{\phi}(\cdot - n) \rangle$$

$$= \sum_k c_k \sum_\ell p_\ell \int_{-\infty}^{\infty} \phi(x - 2k - \ell) \widetilde{\phi}(x - n) dx$$

$$+ \sum_k d_k \sum_\ell q_\ell \int_{-\infty}^{\infty} \phi(x - 2k - \ell) \widetilde{\phi}(x - n) dx$$

$$= \sum_k c_k p_{n-2k} + \sum_k d_k q_{n-2k}.$$

Thus, the sequence $\{x_k\}$ can be recovered from its decomposed components $\{c_k\}$ and $\{d_k\}$.

To summarize, for an input sequence $\mathbf{x} = \{x_k\}$, the **wavelet decomposition algorithm** with **analysis filters** $\widetilde{p}, \widetilde{q}$ is defined by

$$c_n = \frac{1}{2} \sum_k \widetilde{p}_{k-2n}\, x_k, \quad d_n = \frac{1}{2} \sum_k \widetilde{q}_{k-2n}\, x_k, \quad n \in \mathbb{Z}; \tag{8.3.4}$$

and the **wavelet reconstruction algorithm** with **synthesis filters** p, q is given by

$$\widetilde{x}_n = \sum_k p_{n-2k}\, c_k + \sum_k q_{n-2k}\, d_k, \quad n \in \mathbb{Z}. \tag{8.3.5}$$

The outputs $\mathbf{c} = \{c_k\}$ and $\mathbf{d} = \{d_k\}$ are called, respectively, **lowpass output** and **highpass output**, or **approximant** and **details** of $\mathbf{x}$ with the analysis filters $\widetilde{p}, \widetilde{q}$, respectively. Algorithms (8.3.4) and (8.3.5) are also called **discrete wavelet transform (DWT)** and **inverse discrete wavelet transform (IDWT)**, respectively. The reason for introducing the terminology of DWT can be justified as follows.

Set $f^0(x) = f(x)$, $c_k^0 = x_k$ in (8.3.1). Since

$$\mathbb{V}_0 = \mathbb{V}_{-1} + \mathbb{W}_{-1} = \mathbb{V}_{-2} + \mathbb{W}_{-2} + \mathbb{W}_{-1} = \cdots$$
$$= \mathbb{V}_{-J} + \mathbb{W}_{-1} + \cdots + \mathbb{W}_{-J},$$

where J is a positive integer, $f \in \mathbb{V}_0$ can be expanded as

$$f = f^0 = f^{-1} + g^{-1} = f^{-2} + g^{-2} + g^{-1} = \cdots$$
$$= f^{-m} + g^{-1} + \cdots + g^{-m} = \cdots \qquad (8.3.6)$$
$$= f^{-J} + g^{-1} + \cdots + g^{-J},$$

where

$$f^{-j}(x) = \sum_k c_k^{-j} \phi\left(\frac{x}{2^j} - k\right),$$
$$g^{-j}(x) = \sum_k d_k^{-j} \psi\left(\frac{x}{2^j} - k\right).$$

From (8.3.2) and (8.3.4), we see $c_k^{-1} = c_k$, $d_k^{-1} = d_k$. We can obtain, as we derived (8.3.4), that other c_k^{-j} and d_k^{-j} are given iteratively by

$$c_n^{-j} = \frac{1}{2}\sum_k \widetilde{p}_{k-2n} c_k^{1-j}, \quad d_n^{-j} = \frac{1}{2}\sum_k \widetilde{q}_{k-2n} c_k^{1-j}, \quad n \in \mathbb{Z}, \qquad (8.3.7)$$

for $j = 1, 2, \ldots, J$. Now, in view of (8.2.38) and (8.2.32) in the previous section, we see that

$$\langle \phi\left(\frac{x}{2^m} - k\right), \widetilde{\psi}\left(\frac{x}{2^m} - n\right)\rangle = 0;$$
$$\langle \psi\left(\frac{x}{2^j} - k\right), \widetilde{\psi}\left(\frac{x}{2^m} - n\right)\rangle = 2^m \delta_{j-m} \delta_{k-n}$$

for all $k, n \in \mathbb{Z}$ and $j \leq m$. Therefore, it follows from (8.3.6) that

$$\frac{1}{2^m}\langle f(x), \widetilde{\psi}\left(\frac{x}{2^m} - n\right)\rangle = \frac{1}{2^m}\langle f^{-m}(x) + g^{-1}(x) + \cdots + g^{-m}(x), \widetilde{\psi}\left(\frac{x}{2^m} - n\right)\rangle$$
$$= \frac{1}{2^m}\langle \sum_k d_k^{-m} \psi\left(\frac{x}{2^m} - k\right), \widetilde{\psi}\left(\frac{x}{2^m} - n\right)\rangle = d_n^{-m}.$$

That is,

$$d_n^{-m} = \frac{1}{2^m}\int_{-\infty}^{\infty} f(x)\widetilde{\psi}\left(\frac{x - n2^m}{2^m}\right) dx.$$

In other words, according to the definition of the wavelet transform $\left(W_{\widetilde{\psi}}\right)(b, a)$ in (8.1.5) of Sect. 8.1 on p.382, we have

$$d_n^{-m} = \left(W_{\widetilde{\psi}} f\right)(n2^m, 2^m).$$

Of course, if we replace $f(x)$ by $f(2^N x)$ for any positive integer N, then the above derivation yields

$$d_n^j = \left(W_{\widetilde{\psi}} f\right)\left(\frac{n}{2^j}, \frac{1}{2^j}\right), \ n \in \mathbb{Z},$$

for all integers $j < N$. Hence, the wavelet decomposition algorithm (8.3.4) can be applied iteratively j times to compute the wavelet transform $\left(W_{\widetilde{\psi}}\right)(b, a)$ for the time-scale value $(b, a) = \left(\frac{n}{2^j}, \frac{1}{2^j}\right)$ for all integers $n \in \mathbb{Z}$. For this reason, (8.3.4) is called the DWT (algorithm). Furthermore, since (8.3.5) can be applied iteratively j times to recover $\{x_k\}$ (and hence $f(x)$) from $\{c_n^j\}$ and $\{d_n^j\}$, it is called the IDWT (algorithm).

Remark 1 In contrast to the discrete Fourier transform (DFT), which is obtained by discretization of the Fourier coefficients studied in Chaps. 4 and 6, respectively, the discrete wavelet transform is precisely the wavelet transform evaluated at the dyadic time-scale position $\left(\frac{n}{2^j}, \frac{1}{2^j}\right), n \in \mathbb{Z}$, without the need of discretization of the integral $\left(W_{\widetilde{\psi}} f\right)\left(\frac{n}{2^j}, \frac{1}{2^j}\right)$ in (8.1.5) in Sect. 8.1. Therefore, the DWT algorithm (8.3.4) is not only much more efficient than discretization of the integral (8.1.5) in Sect. 8.1, but also yields precise values as opposed to approximate values, of the wavelet transform.   ■

We have shown above that if $\psi, \widetilde{\psi}$ form a pair of biorthogonal wavelets, then the IDWT recovers **x** from its approximant and details. Actually, the condition that $\psi, \widetilde{\psi}$ form a pair of biorthogonal wavelets could be relaxed to the condition that the filter bank $p, q, \widetilde{p}, \widetilde{q}$ are biorthogonal. In the next theorem, we show that a biorthogonal filter bank can be applied to recover the original sequence. To this end, we introduce the *z*-**transform** $X(z)$ of a bi-infinite sequence $\mathbf{x} = \{x_j\}$ defined by

$$X(z) = \sum_{j=-\infty}^{\infty} x_j z^j,$$

where $z \in \mathbb{C}, z \neq 0$. $X(z)$ is only a "symbol", since we are not concerned with its convergence (see the comment on the notion of FIR at the end of our discussion of the two-scale symbol in (8.2.19) on p.396). However, if $x_j = 0$ except for only finitely many terms, then the bi-infinite sum $\sum_{-\infty}^{\infty}$ is a "Laurent polynomial" in z or z^{-1}. Here, we derive $X(z)$ from the definition of the traditional z-transform $\sum_{j=-\infty}^{\infty} x_j z^{-j}$, by using powers z^j instead of z^{-j} (for convenience).

Recall that if a finite sequence $\{g_k\}$ is considered as a lowpass or highpass filter for a scaling function or a wavelet, respectively, then $G(z)$ in (8.2.19) on p.396 denotes

its two-scale symbol (or frequency response), with the multiplicative constant factor of $\frac{1}{2}$:

$$G(z) = \frac{1}{2} \sum_k g_k z^k.$$

Also recall that when $p = \{p_k\}$, $\widetilde{p} = \{\widetilde{p}_k\}$ and $q = \{q_k\}$, $\widetilde{q} = \{\widetilde{q}_k\}$ are considered as the refinement and wavelet sequences, we do not bold-face p, $\widetilde{p}$, q, $\widetilde{q}$ throughout this book (see Remark 1 on p.396).

Theorem 1 | **Biorth. filter bank means perfect recovery of signal** | *If $\widetilde{p}$, $\widetilde{q}$, p and q are biorthogonal, then an input $\mathbf{x}$ can be recovered from its approximant $\mathbf{c}$ and details $\mathbf{d}$ defined by* (8.3.4), *namely : $\widetilde{\mathbf{x}}$ defined by* (8.3.5) *is exactly $\mathbf{x}$.*

Proof Let $C(z)$, $D(z)$, $X(z)$ and $\widetilde{X}(z)$ denote the z-transforms of $\mathbf{c}$, $\mathbf{d}$, $\mathbf{x}$ and $\widetilde{\mathbf{x}}$. Then in the frequency domain, (8.3.4) and (8.3.5) can be written respectively as

$$C(z^2) = \frac{1}{2}\left(\widetilde{P}\left(\frac{1}{z}\right)X(z) + \widetilde{P}\left(-\frac{1}{z}\right)X(-z)\right);$$

$$D(z^2) = \frac{1}{2}\left(\widetilde{Q}\left(\frac{1}{z}\right)X(z) + \widetilde{Q}\left(-\frac{1}{z}\right)X(-z)\right);$$

$$(8.3.8)$$

and

$$\widetilde{X}(z) = 2P(z)C(z^2) + 2Q(z)D(z^2) \tag{8.3.9}$$

(see Exercise 1).

Let $M_{\widetilde{p},\widetilde{q}}$ and $M_{p,q}$ be the **modulation matrices** of $\widetilde{p}$, $\widetilde{q}$ and p, q, respectively, defined by

$$M_{\widetilde{p},\widetilde{q}}(z) = \begin{bmatrix} \widetilde{P}(z) & \widetilde{P}(-z) \\ \widetilde{Q}(z) & \widetilde{Q}(-z) \end{bmatrix}, \quad M_{p,q}(z) = \begin{bmatrix} P(z) & P(-z) \\ Q(z) & Q(-z) \end{bmatrix}.$$

It is easy to see that $\widetilde{p}$, $\widetilde{q}$, p and q are biorthogonal, namely: $\widetilde{P}(z)$, $\widetilde{Q}(z)$, $P(z)$, $Q(z)$ satisfy the biorthogonality conditions (8.2.33)–(8.2.36) on p.401, if and only if

$$M_{\widetilde{p},\widetilde{q}}(z)\left(M_{p,q}\left(\frac{1}{z}\right)\right)^T = I_2, \ z \in \mathbb{C}\backslash\{0\}, \tag{8.3.10}$$

which means that $M_{p,q}\left(\frac{1}{z}\right)^T$ is the inverse matrix of $M_{\widetilde{p},\widetilde{q}}(z)$. Thus, with z replaced by $\frac{1}{z}$, (8.3.10) is equivalent to

$$\left(M_{p,q}(z)\right)^T M_{\widetilde{p},\widetilde{q}}\left(\frac{1}{z}\right) = I_2, \ , z \in \mathbb{C}\backslash\{0\}, \tag{8.3.11}$$

which, in turn, is equivalent to

$$P(z)\widetilde{P}\left(\frac{1}{z}\right) + Q(z)\widetilde{Q}\left(\frac{1}{z}\right) = 1, \quad P(z)\widetilde{P}\left(-\frac{1}{z}\right) + Q(z)\widetilde{Q}\left(-\frac{1}{z}\right) = 0. \quad (8.3.12)$$

Plugging $C(z^2)$ and $D(z^2)$ [as given in (8.3.8)] into (8.3.9), we have

$$\begin{aligned}
\widetilde{X}(z) &= 2\, P(z)\frac{1}{2}\left(\widetilde{P}\left(\frac{1}{z}\right)X(z) + \widetilde{P}\left(-\frac{1}{z}\right)X(-z)\right) \\
&\quad + 2\, Q(z)\frac{1}{2}\left(\widetilde{Q}\left(\frac{1}{z}\right)X(z) + \widetilde{Q}\left(-\frac{1}{z}\right)X(-z)\right) \\
&= \left(P(z)\widetilde{P}\left(\frac{1}{z}\right) + Q(z)\widetilde{Q}\left(\frac{1}{z}\right)\right)X(z) + \left(P(z)\widetilde{P}\left(-\frac{1}{z}\right) + Q(z)\widetilde{Q}\left(-\frac{1}{z}\right)\right)X(-z).
\end{aligned}$$

Thus $\widetilde{X}(z) = X(z)$ if (8.3.12) holds, which is equivalent to saying that $\widetilde{p}, \widetilde{q}, p, q$ form a biorthogonal filter bank. Hence, the biorthogonality of $\widetilde{p}, \widetilde{q}, p$ and q implies that **x** can be recovered from its lowpass and highpass outputs **c**, **d** by the wavelet reconstruction algorithm. ∎

Remark 2 $\boxed{\textbf{Biorthogonality in terms of modulation matrices}}$ From the above proof, we see that $\widetilde{p}, \widetilde{q}, p, q$ form a biorthogonal filter bank if and only if their modulation matrices $M_{\widetilde{p},\widetilde{q}}$, $M_{p,q}$ satisfy (8.3.11). ∎

$\boxed{\textbf{Multi-level wavelet decomposition and reconstruction}}$ The wavelet decomposition algorithm (8.3.4) could be applied to the approximant **c** of an input **x** to get further approximant and details of **x**. Continuing this procedure, we have multiscale (or multi-level) wavelet decomposition and reconstruction algorithms. More precisely, let $\mathbf{c}^{(0)}$ denote the input **x**, that is, $c_k^{(0)} = x_k$. For $J > 1$, the J-level wavelet decomposition algorithm is given by (8.3.7). For biorthogonal $\widetilde{p}(\omega)$, $\widetilde{q}(\omega)$, $p(\omega)$ and $q(\omega)$, **x** is recovered by the corresponding reconstruction algorithm

$$c_n^{1-j} = \sum_k p_{n-2k}c_k^{-j} + \sum_k q_{n-2k}d_k^{-j}, \quad n \in \mathbb{Z}, \quad (8.3.13)$$

for $j = J, J-1, \ldots, 1$. ∎

Example 1 Let $\mathbf{x} = \{x_k\}$ be the sequence with

$$x_k = \begin{cases} 1, & k = 1, 2, \ldots, 16, \\ 0, & \text{elsewhere.} \end{cases}$$

Let $\{p_k\}$, $\{q_k\}$ be the Haar filters with nonzero p_k, q_k given by

$$p_0 = p_1 = 1, \quad q_0 = 1, q_1 = -1.$$

Then, with $c_k^{(0)} = x_k$, the J-level DWT is given by

$$c_n^{-j} = \frac{1}{2}\left(c_{2n}^{1-j} + c_{2n+1}^{1-j}\right), \quad d_n^{-j} = \frac{1}{2}\left(c_{2n}^{1-j} - c_{2n+1}^{1-j}\right), \quad n \in \mathbb{Z},$$

for $j = 1, 2, \ldots, J$. In particular, the approximant c_k^{-1} and the detail d_k^{-1} of $\mathbf{x}$ after 1-level wavelet transform decomposition are given by

$$c_n^{-1} = \begin{cases} \dfrac{1}{2}, & n = 0, 8, \\ 1, & n = 1, 2, \ldots 7, \\ 0, & \text{elsewhere;} \end{cases}$$

and

$$d_n^{-1} = \begin{cases} -\dfrac{1}{2}, & n = 0, \\ \dfrac{1}{2}, & n = 8, \\ 0, & \text{elsewhere.} \end{cases} \qquad \blacksquare$$

Example 2 In this example we show the graphs of the details and approximant of an audio signal after 3-level wavelet decomposition with the D_4 orthogonal filter. The original signal is shown on the top of Fig. 8.4 and the approximant at the bottom of the figure. The details are in the middle of this picture. Observe that after each level of DWT, the details and approximants are essentially halved. $\blacksquare$

$\boxed{\textbf{DWT and orthogonal transformations}}$ For a finite signal, after a periodic extension, the orthogonal DWT can be expressed as a linear transformation of an orthogonal matrix. Let $p = \{p_k\}$ be a QMF supported on $[0, N]$ for some odd N, namely: $p_k = 0$ for $k < 0$ or $k > N$. Let $q = \{q_k\}$ be a corresponding high-pass filter with $q_k = (-1)^k p_{N-k}$, such that q is also supported on $[0, N]$. Let $\{x_0, x_1, \ldots, x_{2L-1}\}$ be a signal with even length, say $2L$, for some $L > N$. Extend this finite signal $2L$-periodically to an infinite sequence, denoted by $\mathbf{x} = \{x_k\}, k \in \mathbb{Z}$. Let $\mathbf{c}$ and $\mathbf{d}$ be the approximant and detail of $\mathbf{x}$, respectively, obtained by (8.3.4) with analysis filters p and q. Then both $\mathbf{c}$ and $\mathbf{d}$ are L-periodic (see Exercise 8). Furthermore, we have

$$\begin{bmatrix} c_0 \\ \vdots \\ c_{L-1} \end{bmatrix} = \frac{1}{2}\mathcal{P}\begin{bmatrix} x_0 \\ \vdots \\ x_{2L-1} \end{bmatrix}, \quad \begin{bmatrix} d_0 \\ \vdots \\ d_{L-1} \end{bmatrix} = \frac{1}{2}\mathcal{Q}\begin{bmatrix} x_0 \\ \vdots \\ x_{2L-1} \end{bmatrix}, \tag{8.3.14}$$

where $\mathcal{P}$ and $\mathcal{Q}$ are $L \times (2L)$ matrices given by
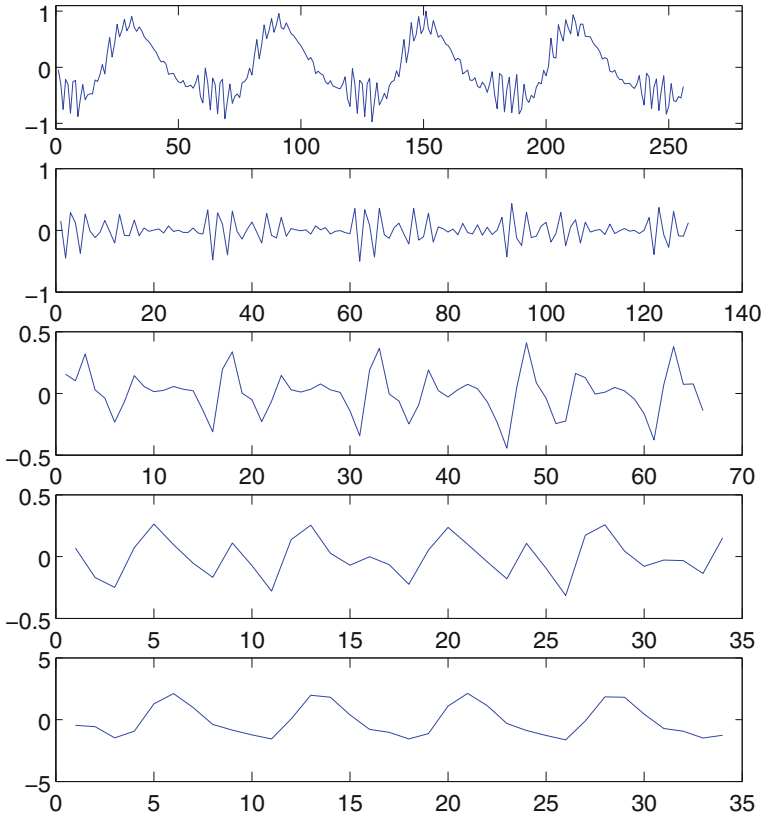
Fig. 8.4. From *top* to *bottom*: original signal, details after 1-, 2-, 3-level DWT, and the approximant

$$
\mathcal{P} =
\begin{bmatrix}
p_0 & p_1 & p_2 & p_3 & \cdots & p_N & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\
0 & 0 & p_0 & p_1 & \cdots & p_{N-2} & p_{N-1} & p_N & \cdots & 0 & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
p_4 & p_5 & p_6 & p_7 & \cdots & 0 & 0 & 0 & \cdots & p_0 & p_1 & p_2 & p_3 \\
p_2 & p_3 & p_4 & p_5 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & p_0 & p_1
\end{bmatrix},
\tag{8.3.15}
$$

$$
\mathcal{Q} =
\begin{bmatrix}
q_0 & q_1 & q_2 & q_3 & \cdots & q_N & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\
0 & 0 & q_0 & q_1 & \cdots & q_{N-2} & q_{N-1} & q_N & \cdots & 0 & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
q_4 & q_5 & q_6 & q_7 & \cdots & 0 & 0 & 0 & \cdots & q_0 & q_1 & q_2 & q_3 \\
q_2 & q_3 & q_4 & q_5 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & q_0 & q_1
\end{bmatrix}.
\tag{8.3.16}
$$

Let $\widetilde{\mathbf{x}}$ be the sequence obtained by IDWT (8.3.5). Then it can be shown, by using the L-periodicity of $\mathbf{c}$ and $\mathbf{d}$, that $\widetilde{\mathbf{x}}$ is $2L$-periodic (see Exercise 10), and that

$$
\begin{bmatrix} \widetilde{x}_0 \\ \vdots \\ \widetilde{x}_{2L-1} \end{bmatrix} = \mathcal{P}^T \begin{bmatrix} c_0 \\ \vdots \\ c_{L-1} \end{bmatrix} + \mathcal{Q}^T \begin{bmatrix} d_0 \\ \vdots \\ d_{L-1} \end{bmatrix}. \tag{8.3.17}
$$

Thus, from (8.3.14) and (8.3.17), we have

$$
\begin{bmatrix} \widetilde{x}_0 \\ \vdots \\ \widetilde{x}_{2L-1} \end{bmatrix} = \mathcal{P}^T \frac{1}{2} \mathcal{P} \begin{bmatrix} x_0 \\ \vdots \\ x_{2L-1} \end{bmatrix} + \mathcal{Q}^T \frac{1}{2} \mathcal{Q} \begin{bmatrix} x_0 \\ \vdots \\ x_{2L-1} \end{bmatrix} = \frac{1}{2} \left(\mathcal{P}^T \mathcal{P} + \mathcal{Q}^T \mathcal{Q} \right) \begin{bmatrix} x_0 \\ \vdots \\ x_{2L-1} \end{bmatrix}.
$$

Hence, if p and q are orthogonal, then $\widetilde{\mathbf{x}} = \mathbf{x}$, and we conclude that

$$
\mathcal{P}^T \mathcal{P} + \mathcal{Q}^T \mathcal{Q} = 2I,
$$

which leads to the following theorem.

Theorem 2 | **DWT and orthogonal transformations** | *Let p and q be orthogonal filters supported on $[0, N]$. Define*

$$
\mathcal{W} = \begin{bmatrix} \mathcal{P} \\ \mathcal{Q} \end{bmatrix}, \tag{8.3.18}
$$

where $\mathcal{P}$ and $\mathcal{Q}$ are defined by (8.3.15) and (8.3.16). Then $\frac{1}{\sqrt{2}} \mathcal{W}$ is an orthogonal matrix, namely:

$$
\mathcal{W}^T \mathcal{W} = 2I. \tag{8.3.19}
$$

In fact, (8.3.19) can be derived directly from the orthogonality of p, q given in (8.2.25)–(8.2.27) of the previous section on p.399 (see Exercises 11–14). Denote

$$
\mathbf{x}_{2L} = [x_0, \ldots, x_{2L-1}]^T, \ \mathbf{c}_L = [c_0, \ldots, c_{L-1}]^T,
$$
$$
\mathbf{d}_L = [d_0, \ldots, d_{L-1}]^T, \ \widetilde{\mathbf{x}}_{2L} = [\widetilde{x}_0, \ldots, \widetilde{x}_{2L-1}]^T.
$$

Then (8.3.14) and (8.3.17) can be written as

$$
\begin{bmatrix} \mathbf{c}_L \\ \mathbf{d}_L \end{bmatrix} = \frac{1}{2} \mathcal{W} \mathbf{x}_{2L}, \quad \widetilde{\mathbf{x}}_{2L} = \mathcal{W}^T \begin{bmatrix} \mathbf{c}_L \\ \mathbf{d}_L \end{bmatrix}.
$$

Thus, DWT and IDWT can be expressed as orthogonal transformations. ∎

DWT and lifting scheme   DWT and IDWT can be implemented as lifting-scheme algorithms. This topic will be discussed in some detail in Sect. 9.4 of the next chapter.

Haar filters For example, for the unnormalized Haar filter bank $\widetilde{L}(z) = \frac{1}{2}$ $(1 + z)$, $\widetilde{H}(z) = \frac{1}{2}(1 - z)$ and $L(z) = 1 + z$, $H(z) = 1 - z$, the corresponding (modified) forward and backward lifting algorithms for input $\mathbf{x} = (x_k)$ are given as follows.

> **Forward lifting (wavelet decomposition) algorithm**:
>
> **Splitting**: $\mathbf{c} = (x_{2k})$, $\mathbf{d} = (x_{2k+1})$;
>
> **Step1**. $\mathbf{c}^{(1)} = \dfrac{1}{2}(\mathbf{c} + \mathbf{d})$;
>
> **Step2**. $\mathbf{d}^{(1)} = \mathbf{d} - \mathbf{c}^{(1)}$.
>
> $\mathbf{c}^{(1)}$ and $\mathbf{d}^{(1)}$ are the lowpass and highpass outputs.

> **Backward lifting (wavelet reconstruction) algorithm**:
>
> **Step1**. $\mathbf{d} = \mathbf{d}^{(1)} + \mathbf{c}^{(1)}$;
>
> **Step2**. $\mathbf{c} = 2\mathbf{c}^{(1)} - \mathbf{d}$;
>
> **Combining**: $\mathbf{x} = (\ldots, c_k, d_k, \ldots)$.
>
> $\mathbf{x}$ is recovered from its lowpass and highpass outputs $\mathbf{c}^{(1)}$ and $\mathbf{d}^{(1)}$.

See Fig. 8.5 for the algorithms.

5/3-tap biorthogonal filters As another example, DWT and IDWT with 5/3-biorthogonal filter bank considered in Example 5 on p.402 can be written as (modified) forward and backward lifting for input $\mathbf{x}$:

> **Forward lifting (wavelet decomposition) algorithm**:
>
> **Splitting**: $\mathbf{c} = (x_{2k})$, $\mathbf{d} = (x_{2k+1})$;
>
> **Step1**. $c_k^{(1)} = c_k - \dfrac{1}{2}(d_k + d_{k-1})$;
>
> **Step2**. $d_k^{(1)} = d_k + \dfrac{1}{4}\left(c_k^{(1)} + c_{k+1}^{(1)}\right)$;
>
> **Step3**. $c_k^{(2)} = d_k^{(1)}$, $d_k^{(2)} = \dfrac{1}{2}c_k^{(1)}$.
>
> $\mathbf{c}^{(2)}$ and $\mathbf{d}^{(2)}$ are the lowpass and highpass outputs.

> **Backward lifting (wavelet reconstruction) algorithm**:
>
> **Step1**. $c_k^{(1)} = 2d_k^{(2)}$, $d_k^{(1)} = c_k^{(2)}$;
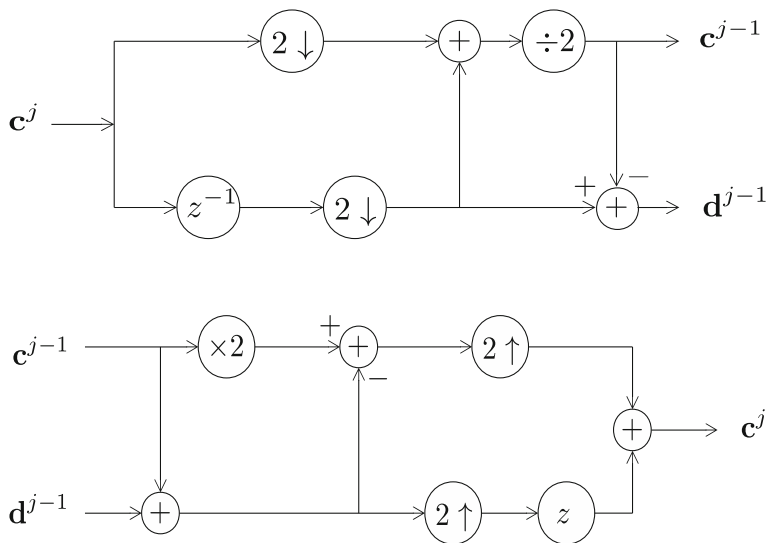
Fig. 8.5. From *top* to *bottom*: (modified) forward and backward lifting algorithms of Haar filters

Step2. $d_k = d_k^{(1)} - \dfrac{1}{4}\left(c_k^{(1)} + c_{k+1}^{(1)}\right)$;

Step3. $c_k = c_k^{(1)} + \dfrac{1}{2}(d_k + d_{k-1})$;

Combining: $\mathbf{x} = (\ldots, c_k, d_k, \ldots).$

$\mathbf{x}$ is recovered from its lowpass and highpass outputs $\mathbf{c}^{(2)}$ and $\mathbf{d}^{(1)}$.

See Fig. 8.6 for the algorithms.   ∎

Two-dimensional DWT The MRA architecture can be used to construct compactly supported wavelets of higher dimensions. Here we provide the tensor-product wavelets (also called separable wavelets) on $\mathbb{R}^2$ and the associated DWT and IDWT.

Suppose ϕ is a compactly supported scaling function with an FIR lowpass filter $p = \{p_k\}$, and ψ is an associated compactly supported wavelet with a highpass filter $q = \{q_k\}$. Recall that the nested sequence of spaces

$$\mathbb{V}_j = \overline{\text{span}}\{\phi(2^j x - k) : \ k \in \mathbb{Z}\}, \quad -\infty < j < \infty,$$

constitutes an orthogonal MRA, and $\mathbb{V}_{j+1} = \mathbb{V}_j \oplus^\perp \mathbb{W}_j$ (namely, $\mathbb{V}_{j+1} = \mathbb{V}_j + \mathbb{W}_j$ and $\mathbb{V}_j \perp \mathbb{W}_j$), where $\mathbb{W}_j = \overline{\text{span}}\{\psi(2^j x - k) : \ k \in \mathbb{Z}\}$.

Define

$$\mathcal{V}_j = \mathbb{V}_j \otimes \mathbb{V}_j = \{f(x)g(y) : \ f(x) \in \mathbb{V}_j, g(y) \in \mathbb{V}_j\}.$$
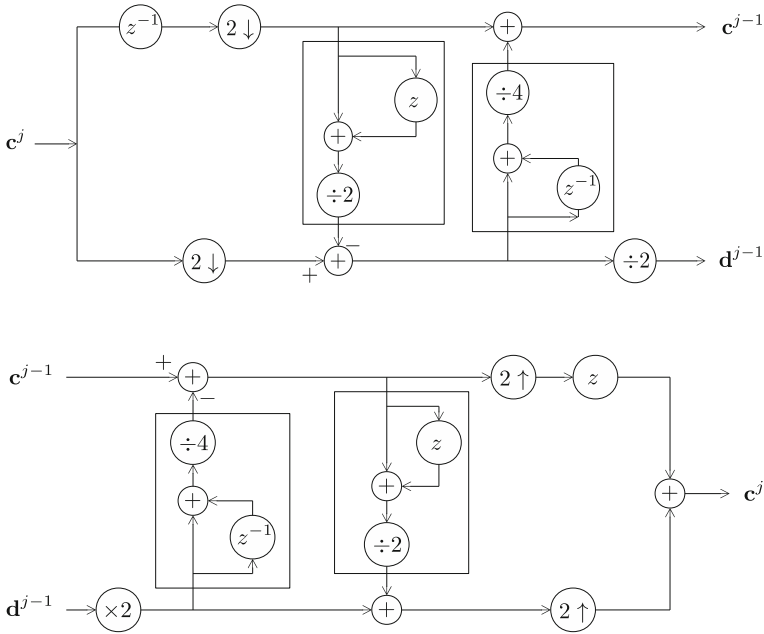
Fig. 8.6. From *top* to *bottom*: (modified) forward and backward lifting algorithms of 5/3-tap biorthogonal filters

Then $\{\mathcal{V}_j\}$, $j \in \mathbb{Z}$, form a nested sequence of subspaces of $L_2(\mathbb{R}^2)$ with the density property, namely;

$$\{0\} \leftarrow \cdots \subset \mathcal{V}_{-1} \subset \mathcal{V}_0 \subset \mathcal{V}_1 \subset \mathcal{V}_2 \subset \cdots \rightarrow L_2(\mathbb{R}^2).$$

In addition, with the notations

$$\mathcal{W}_j^{(1)} = \mathbb{W}_j \otimes \mathbb{V}_j, \ \mathcal{W}_j^{(2)} = \mathbb{V}_j \otimes \mathbb{W}_j, \ \mathcal{W}_j^{(3)} = \mathbb{W}_j \otimes \mathbb{W}_j,$$

we have

$$\mathcal{V}_{j+1} = \mathcal{V}_j \oplus^\perp \mathcal{W}_j^{(1)} \oplus^\perp \mathcal{W}_j^{(2)} \oplus^\perp \mathcal{W}_j^{(3)}.$$

Let us now introduce the 2-dimensional separable scaling function and wavelets

$$\Phi(x, y) = \phi(x)\phi(y), \quad \Psi^{(1)}(x, y) = \psi(x)\phi(y),$$
$$\Psi^{(2)}(x, y) = \phi(x)\psi(y), \quad \Psi^{(3)}(x, y) = \psi(x)\psi(y),$$

and for $f(x, y)$ defined in $\mathbb{R}^2$, denote

$$f_{j,k_1,k_2}(x, y) = 2^j f(2^j x - k_1, 2^j y - k_2), \ j, k_1, k_2 \in \mathbb{Z}.$$

Then

$$\left\{\Phi_{j,k_1,k_2} : k_1, k_2 \in \mathbb{Z}\right\}, \quad \left\{\Psi^{(1)}_{j,k_1,k_2} : k_1, k_2 \in \mathbb{Z}\right\},$$
$$\left\{\Psi^{(2)}_{j,k_1,k_2} : k_1, k_2 \in \mathbb{Z}\right\}, \quad \left\{\Psi^{(3)}_{j,k_1,k_2} : k_1, k_2 \in \mathbb{Z}\right\},$$

are orthonormal bases of $\mathcal{V}_j, \mathcal{W}^{(1)}_j, \mathcal{W}^{(2)}_j, \mathcal{W}^{(3)}_j$, respectively. This, together with the fact that

$$\oplus^{\perp}_{j \in \mathbb{Z}} \left(\oplus^{\perp} \mathcal{W}^{(1)}_j \oplus^{\perp} \mathcal{W}^{(2)}_j \oplus^{\perp} \mathcal{W}^{(3)}_j \right) = L_2(\mathbb{R}^2),$$

implies that

$$\left\{\Psi^{(1)}_{j,k_1,k_2}, \Psi^{(2)}_{j,k_1,k_2}, \Psi^{(3)}_{j,k_1,k_2} : j, k_1, k_2 \in \mathbb{Z}\right\}$$

is an orthonormal basis of $L_2(\mathbb{R}^2)$.

For a sequence $\mathbf{c}^{(0)} = \left\{c^{(0)}_{k_1,k_2} : k_1, k_2 \in \mathbb{Z}\right\}$, the J-level 2-dimensional (tensor-product) DWT is given by

$$\begin{cases} c^{-j}_{n_1,n_2} = \dfrac{1}{4} \displaystyle\sum_{k_1,k_2} p_{k_1-2n_1} p_{k_2-2n_2} c^{1-j}_{k_1,k_2}, \\[2mm] d^{-j,1}_{n_1,n_2} = \dfrac{1}{4} \displaystyle\sum_{k_1,k_2} q_{k_1-2n_1} p_{k_2-2n_2} c^{1-j}_{k_1,k_2}, \\[2mm] d^{-j,2}_{n_1,n_2} = \dfrac{1}{4} \displaystyle\sum_{k_1,k_2} p_{k_1-2n_1} q_{k_2-2n_2} c^{1-j}_{k_1,k_2}, \\[2mm] d^{-j,3}_{n_1,n_2} = \dfrac{1}{4} \displaystyle\sum_{k_1,k_2} q_{k_1-2n_1} q_{k_2-2n_2} c^{1-j}_{k_1,k_2}, \quad n_1, n_2 \in \mathbb{Z}, \end{cases} \tag{8.3.20}$$

for $j = 1, 2, \ldots, J$; and the 2-dimensional (tensor-product) IDWT is given by

$$c^{1-j}_{n_1,n_2} = \sum_{k_1,k_2} p_{n_1-2k_1} p_{n_2-2k_2} c^{-j}_{k_1,k_2} + \sum_{k_1,k_2} q_{n_1-2k_1} p_{n_2-2k_2} d^{-j,1}_{k_1,k_2} \tag{8.3.21}$$
$$+ \sum_{k_1,k_2} p_{n_1-2k_1} q_{n_2-2k_2} d^{-j,2}_{k_1,k_2} + \sum_{k_1,k_2} q_{n_1-2k_1} q_{n_2-2k_2} d^{-j,3}_{k_1,k_2},$$

for $j = J, J-1, \ldots, 1$. Each $\mathbf{c}^{-j}$ is called a (j-level) approximant of $\mathbf{c}^{(0)}$, and $\mathbf{d}^{-j,1}, \mathbf{d}^{-j,2}, \mathbf{d}^{-j,3}$ are called (j-level) details of $\mathbf{c}^{(0)}$.

For a biorthogonal filter bank $\widetilde{p} = \{\widetilde{p}_k\}, \widetilde{q} = \{\widetilde{q}_k\}, p = \{p_k\}, q = \{q_k\}$, the 2-dimensional DWT is given by (8.3.20), with p_k and q_k being replaced by $\widetilde{p}_k$ and $\widetilde{q}_k$, respectively, and the IDWT is given by (8.3.21).

<div align="center">

Exercises

</div>

Exercise 1 Show that (8.3.4) and (8.3.5) are equivalent to (8.3.8) and (8.3.9), respectively.

Exercise 2 Derive (8.3.3).

Exercise 3 Let $p = \{p_k\}, q = \{q_k\}$ be the Haar filters with $p_0 = p_1 = 1$, $q_0 = 1, q_1 = -1$ and $p_k = 0, q_k = 0$ for $k < 0$ or $k > 1$. Formulate the modulation matrix $M_{p,q}(z)$ of p and q, and verify that

$$M_{p,q}(z) \left(M_{p,q}\left(\frac{1}{z}\right) \right)^T = I_2, \ z \in \mathbb{C}\backslash\{0\}.$$

Exercise 4 Repeat Exercise 3 with p, q being the D_4-filters given in Example 4 of the previous section on p.399.

Exercise 5 Let $\widetilde{p}, \widetilde{q}, p, q$ be the 5/3-tap biorthogonal filters given in Example 5 of the previous section on p.402. Verify the validity of (8.3.10) for this filter bank.

Exercise 6 Let $\mathbf{x} = \{x_k\}$ be the sequence with $x_k = 1$ for $1 \leq k \leq 8$ and $x_k = 0$ for $k < 1$ or $k > 8$. Determine the approximant $\mathbf{c} = \{c_k\}$ and the details $\mathbf{d} = \{d_k\}$ of $\mathbf{x}$ after 1-level wavelet decomposition with the D_4-filters p and q given on p.400.

Exercise 7 Repeat Exercise 6 with the D_4-filters p, q replaced by $\widetilde{p}, \widetilde{q}$ of 5/3-tap biorthogonal filters given on p.402.

Exercise 8 Let $\mathbf{x} = \{x_k\}$ be a sequence with period $2L$ for some positive integer L, namely: $x_{k+2L} = x_k, k \in \mathbb{Z}$. Show that $\mathbf{y} = \{y_k\}$, defined by

$$y_n = \frac{1}{2} \sum_k h_{k-2n} x_k,$$

for some constants h_k, is L-periodic.

Exercise 9 Let $\widetilde{\mathbf{x}}$ be a sequence defined by (8.3.5) for some sequences $\mathbf{c}$ and $\mathbf{d}$ for some p_k, q_k. Show that if $\mathbf{c}$ and $\mathbf{d}$ are L-periodic, then $\widetilde{\mathbf{x}}$ is $2L$-periodic.

Exercise 10 Derive the formula (8.3.14).

Exercise 11 Derive the formula (8.3.17).

Exercise 12 Let $p = \{p_k\}$ and $q = \{q_k\}$ be two FIR filters, and let $P(z)$ and $Q(z)$ be their two-scale symbols. Show that (8.2.25), (8.2.26), (8.2.27) of the previous section on p.399 are, respectively, equivalent to

(a) $\displaystyle\sum_k p_k\, p_{k-2n} = 2\delta_n, \quad n \in \mathbb{Z};$

(b) $\displaystyle\sum_k q_k q_{k-2n} = 2\delta_n, \quad n \in \mathbb{Z};$

(c) $\displaystyle\sum_k p_k q_{k-2n} = 0, \quad n \in \mathbb{Z}.$

Exercise 13 Let $\mathcal{P}$ and $\mathcal{Q}$ be the $L \times (2L)$ matrices defined by (8.3.15) and (8.3.16) for p and q supported on N with $N < L$. As a continuation of Exercise 12, show that (a), (b), (c) in Exercise 12 are respectively equivalent to

(i) $\mathcal{P}\mathcal{P}^T = 2I_L;$
(ii) $\mathcal{Q}\mathcal{Q}^T = 2I_L;$
(iii) $\mathcal{P}\mathcal{Q}^T = 0.$

Exercise 14 As a continuation of Exercises 12 and 13, show that p and q are orthogonal (i.e. they satisfy (8.2.25), (8.2.26) and (8.2.27) on p.399), if and only if the matrix $\mathcal{W}$, defined by (8.3.18), satisfies (8.3.19).
    Hint: (8.3.19) is equivalent to $\mathcal{W}\mathcal{W}^T = 2I$.

Exercise 15 Derive the formula (8.3.20).

Exercise 16 Derive the formula (8.3.21).

8.4 Perfect-Reconstruction Filter Banks

Recall that the "integral convolution" of two functions $f, g \in (L^2 \cap L^1)(\mathbb{R})$, defined by

$$(f * g)(x) = \int_{-\infty}^{\infty} f(t)g(x-t)dt = \int_{-\infty}^{\infty} f(x-t)g(t)dt,$$

possesses the Fourier transform property

$$\widehat{(f * g)}(\omega) = \widehat{f}(\omega)\widehat{g}(\omega).$$

This important property extends to the z-transform of bi-infinite sequences in $\ell^2 = \ell^2(\mathbb{Z})$. Recall that the z-transform of a sequence $\mathbf{x} = \{x_j\} \in \ell^2$ is defined by

$$X(z) = \sum_{j=-\infty}^{\infty} x_j z^j,$$

and the "**discrete convolution**" of two sequences $\mathbf{x}, \mathbf{y} \in \ell^2$ by

$$(\mathbf{x} * \mathbf{y})_k = \sum_{j=-\infty}^{\infty} x_j y_{k-j} = \sum_{j=-\infty}^{\infty} x_{k-j} y_j.$$

Then it can be shown that the z-transform of the discrete convolution $\mathbf{w} = \mathbf{x} * \mathbf{y}$ of $\mathbf{x}$ and $\mathbf{y}$ is given by

$$W(z) = X(z)Y(z). \tag{8.4.1}$$

Indeed, by a change of summation indices (from $j - k$ to n below), we have

$$
\begin{aligned}
W(z) &= \sum_{j=-\infty}^{\infty} w_j \, z^j = \sum_{j=-\infty}^{\infty} (\mathbf{x} * \mathbf{y})_j \, z^j \\
&= \sum_{j=-\infty}^{\infty} \left(\sum_k x_k \, y_{j-k} \right) z^j = \sum_{k=-\infty}^{\infty} x_k \left(\sum_j y_{j-k} \, z^{j-k} \right) z^k \\
&= \sum_{k=-\infty}^{\infty} x_k \left(\sum_n y_n \, z^n \right) z^k = \left(\sum_{k=-\infty}^{\infty} x_k \, z^k \right) \left(\sum_{n=-\infty}^{\infty} y_n \, z^n \right) \\
&= X(z)Y(z).
\end{aligned}
$$

Example 1 Let $\mathbf{x} = \{x_j\}$ be the sequence defined by $x_0 = 1, x_1 = 1$ and $x_j = 0, j \neq 0, 1$. Then $X(z) = 1 + z$ and

$$
(\mathbf{x} * \mathbf{x})_k = \sum_j x_j \, x_{k-j} = \sum_{j=0}^{1} x_j \, x_{k-j} = \sum_{j=0}^{1} x_{k-j}
$$

$$
= x_k + x_{k-1} = \begin{cases} 1, & \text{if } k = 0, 2, \\ 2, & \text{if } k = 1, \\ 0, & \text{otherwise}, \end{cases}
$$

so that the z-transform of $\mathbf{x} * \mathbf{x}$ is given by $1 + 2z + z^2 = (1 + z)^2 = (X(z))^2$. This verifies the property in (8.4.1), namely: "the z-transform" of the convolution of $\mathbf{x}$ with itself is the square of the z-transform of $\mathbf{x}$.   ∎

In the following, we introduce the very general notion of FIR lowpass and highpass filters. A finite impulse response (FIR) **lowpass filter** $\boldsymbol{\ell} = \{\ell_j\}$ is a sequence of real numbers ℓ_j that satisfies

(i) $\ell_j = 0$ except for finitely many $j \in \mathbb{Z}$,
(ii) $\sum_j \ell_j = c_0$ for a constant $c_0 > 0$, and
(iii) $\sum_j (-1)^j \ell_j = 0$;

and an FIR **highpass filter** $\mathbf{h} = \{h_j\}$ is a sequence of real numbers h_j that satisfies

(i) $h_j = 0$ except for finitely many $j \in \mathbb{Z}$,
(ii) $\sum_j h_j = 0$, and
(iii) $\sum_j (-1)^j h_j = c_0$.

Let $L(z)$ and $H(z)$ be the z-transforms of $\boldsymbol{\ell}$ and $\mathbf{h}$, respectively. Then it is clear from the above definitions of $\boldsymbol{\ell}$ and $\mathbf{h}$, that

Fig. 8.7. Ideal lowpass filter (on *left*) and highpass filter (on *right*)

$$L(1) = c_0, \quad L(-1) = 0;$$

and

$$H(1) = 0, \quad H(-1) = c_0.$$

For $z = e^{-i\omega}$, the two trigonometric polynomials $L(e^{-i\omega})$ and $H(e^{-i\omega})$, as functions of ω, are also called the frequency responses of ℓ and $\mathbf{h}$. We remark that ideal lowpass and highpass filters, with frequency responses shown in Fig. 8.7, are not FIR filters, since the Fourier series on the interval $(0, 2\pi)$ of the characteristic function of $(0, \pi)$, or of $[\pi, 2\pi)$, has infinitely many non-zero coefficients (see Sect. 6.1 of Chap. 6).

The implementation of lowpass and highpass filtering, by using some FIR lowpass and highpass filters $\mathbf{s} = \{s_j\}$ and $\mathbf{r} = \{r_j\}$, respectively, can be represented as follows:

$$\mathbf{x} : (\mathbf{x})_j = x_j \quad \begin{matrix} \nearrow \boxed{*\mathbf{s}} \longrightarrow (\mathbf{x}^L)_k = \sum_j s_j\, x_{k-j}, & \boxed{\textbf{Lowpass output}} \\ \\ \searrow \boxed{*\mathbf{r}} \longrightarrow (\mathbf{x}^H)_k = \sum_j r_j\, x_{k-j}, & \boxed{\textbf{Highpass output}} \end{matrix}$$

where the lowpass and highpass outputs $\mathbf{x}^L$ and $\mathbf{x}^H$ are the low-frequency and high-frequency components of the input $\mathbf{x}$, and they are obtained by taking discrete convolutions with the "finite" sequences $\{s_j\}$ and $\{r_j\}$, which are therefore finite sums.

Example 2 Let $\mathbf{s} = \{s_j\}$ and $\mathbf{r} = \{r_j\}$ be the lowpass and highpass filters defined by $s_{-1} = s_0 = \frac{1}{2}$, $s_j = 0$, $j \neq -1, 0$, and $r_{-1} = -\frac{1}{2}$, $r_0 = \frac{1}{2}$, $r_j = 0$, $j \neq -1, 0$. Then the lowpass output $\mathbf{x}^L$ and highpass output $\mathbf{x}^H$ of the input sequence $\mathbf{x} = \{x_j\}$, obtained by using the filters $\mathbf{s}$ and $\mathbf{r}$, are given by

$$x_k^L = \frac{1}{2}(x_k + x_{k+1}), \quad x_k^H = \frac{1}{2}(x_k - x_{k+1});$$

that is,

$$\mathbf{x}^L = \frac{1}{2} \begin{bmatrix} \vdots \\ x_{-1} + x_0 \\ x_0 + x_1 \\ x_1 + x_2 \\ x_2 + x_3 \\ \vdots \end{bmatrix}, \quad \mathbf{x}^H = \frac{1}{2} \begin{bmatrix} \vdots \\ x_{-1} - x_0 \\ x_0 - x_1 \\ x_1 - x_2 \\ x_2 - x_3 \\ \vdots \end{bmatrix}. \qquad \blacksquare$$

Observe that for a "finite" input sequence $\mathbf{x}$ (i.e. a bi-infinite sequence with finite index-support), the lengths (i.e. the smallest index-support) of the output sequences $\mathbf{x}^L$ and $\mathbf{x}^H$ are equal, and that this length is the same as the length of the input sequence $\mathbf{x}$. So, at least intuitively, "half" of the contents of $\mathbf{x}^L$ and "half" of those of $\mathbf{x}^H$ could be deleted without loss of data information. To accomplish this, we will next introduce the "downsampling operator" D to drop "half" of the lowpass and highpass output sequence components; namely, to replace $\mathbf{x}^L$ and $\mathbf{x}^H$ by $\mathbf{x}^{LD}$ and $\mathbf{x}^{HD}$ that have half of the contents of $\mathbf{x}^L$ and $\mathbf{x}^H$, respectively. We will also introduce the "upsampling operator" U for recovering the input sequence $\mathbf{x}$ from the downsampled outputs $\mathbf{x}^{LD}$ and $\mathbf{x}^{HD}$.

Definition 1 $\boxed{\textbf{Downsampling and upsampling operations}}$ *The operator* $D : \ell^2 (\mathbb{Z}) \to \ell^2(\mathbb{Z})$, *defined by dropping the odd-indexed terms; that is,*

$$D(\dots, x_{-2}, x_{-1}, x_0, x_1, x_2, \dots) = (\dots, x_{-2}, x_0, x_2, \dots),$$

is called the downsampling operator. The operator $U : \ell^2 (\mathbb{Z}) \to \ell^2 (\mathbb{Z})$, *defined by inserting a zero in-between every two consecutive terms; that is,*

$$U(\dots, y_{-1}, y_0, y_1, \dots) = (\dots, 0, y_{-1}, 0, y_0, 0, y_1, 0, \dots),$$

is called the upsampling operator. More precisely, the downsampling and upsampling operators, D and U, are defined by

$$\mathbf{x} \xrightarrow{D} \mathbf{x}^D, \quad \left(\mathbf{x}^D\right)_k = x_{2k}, \ k \in \mathbb{Z};$$

$$\mathbf{y} \xrightarrow{U} U\mathbf{y}, \quad (U\mathbf{y})_{2j} = y_j, \ (U\mathbf{y})_{2j+1} = 0, \ j \in \mathbb{Z},$$

respectively.

The operations of D and U are represented as follows:

$$\mathbf{x} \to \boxed{2\downarrow} \to \mathbf{x}^D;$$

$$\mathbf{y} \to \boxed{2\uparrow} \to U\mathbf{y}.$$

Example 3 Let $\mathbf{x}^L$ and $\mathbf{x}^H$ be the lowpass and highpass outputs of the input sequence $\mathbf{x}$, obtained by applying the filters $\mathbf{s}$ and $\mathbf{r}$ given in Example 2. Then the downsampled sequences $\mathbf{x}^{LD}$ and $\mathbf{x}^{HD}$, obtained by applying the downsampling operator D to $\mathbf{x}^L$ and $\mathbf{x}^H$, respectively, are given by

$$
\mathbf{x}^{LD} = \frac{1}{2}
\begin{bmatrix}
\vdots \\
x_{-2} + x_{-1} \\
x_0 + x_1 \\
x_2 + x_3 \\
\vdots
\end{bmatrix}
, \quad
\mathbf{x}^{HD} = \frac{1}{2}
\begin{bmatrix}
\vdots \\
x_{-2} - x_{-1} \\
x_0 - x_1 \\
x_2 - x_3 \\
\vdots
\end{bmatrix}
.
$$

Next, by applying the upsampling operator U to $\mathbf{x}^{LD}$ and $\mathbf{x}^{HD}$, we obtain

$$
U\mathbf{x}^{LD} = \frac{1}{2}
\begin{bmatrix}
\vdots \\
0 \\
x_{-2} + x_{-1} \\
0 \\
x_0 + x_1 \\
0 \\
x_2 + x_3 \\
0 \\
\vdots
\end{bmatrix}
; \quad
U\mathbf{x}^{HD} = \frac{1}{2}
\begin{bmatrix}
\vdots \\
0 \\
x_{-2} - x_{-1} \\
0 \\
x_0 - x_1 \\
0 \\
x_2 - x_3 \\
0 \\
\vdots
\end{bmatrix}
. \qquad \blacksquare
$$

Example 4 Let $\{\boldsymbol{\ell}, \mathbf{h}\}$ be a lowpass-highpass filter pair, with $\boldsymbol{\ell} = \{\ell_j\}$ and $\mathbf{h} = \{h_j\}$ defined by

$$
\begin{cases}
\ell_0 = \ell_1 = 1, \ \ell_j = 0, \ j \neq 0, 1; \\
h_0 = 1, h_1 = -1, \ h_j = 0, \ j \neq 0, 1,
\end{cases}
\tag{8.4.2}
$$

then for any input sequence $\mathbf{y}$, its lowpass and highpass outputs $\mathbf{y} * \boldsymbol{\ell}$ and $\mathbf{y} * \mathbf{h}$, obtained by applying the filter pair $\{\boldsymbol{\ell}, \mathbf{h}\}$, are given by

$$
(\mathbf{y} * \boldsymbol{\ell})_j = y_j + y_{j-1}, \ (\mathbf{y} * \mathbf{h})_j = y_j - y_{j-1},
$$

by following the same argument as given in Example 2. Thus, as a continuation of Example 3, we observe that

$$
\left(U\mathbf{x}^{LD}\right) * \boldsymbol{\ell} = \frac{1}{2}
\begin{bmatrix}
\vdots \\
x_{-2} + x_{-1} \\
x_{-2} + x_{-1} \\
x_0 + x_1 \\
x_0 + x_1 \\
x_2 + x_3 \\
x_2 + x_3 \\
\vdots
\end{bmatrix},
\quad
\left(U\mathbf{x}^{LD}\right) * \mathbf{h} = \frac{1}{2}
\begin{bmatrix}
\vdots \\
x_{-2} - x_{-1} \\
x_{-1} - x_{-2} \\
x_0 - x_1 \\
x_1 - x_0 \\
x_2 - x_3 \\
x_3 - x_2 \\
\vdots
\end{bmatrix},
$$

so that, by summing the above two results, we have

$$
\left(U\mathbf{x}^{LD}\right) * \boldsymbol{\ell} + \left(U\mathbf{x}^{LD}\right) * \mathbf{h} = \mathbf{x}. \qquad\qquad \blacksquare
$$

We remark that filters $\mathbf{s}, \mathbf{r}, \boldsymbol{\ell}, \mathbf{h}$ in Examples 2, 3 and 4 satisfy

$$
s_j = \frac{1}{2}\ell_{-j}, \ \ r_j = \frac{1}{2}h_{-j}, \ \ j \in \mathbb{Z},
$$

and that $\{\boldsymbol{\ell}, \mathbf{h}\}$ is precisely the Haar filter pair, as introduced earlier.

In what follows, $\mathbf{s}$ and $\mathbf{r}$ are said to be obtained by "time-reversal" of the filters $\frac{1}{2}\boldsymbol{\ell}$ and $\frac{1}{2}\mathbf{h}$, respectively (and for convenience, they are also called "time-reverses" of $\frac{1}{2}\boldsymbol{\ell}$ and $\frac{1}{2}\mathbf{h}$). In general, the **time-reverse** of a (bi-infinite) sequence $\mathbf{g} = \{g_k\}$ is defined by $\mathbf{g}^- = \{g_k^-\}$, where

$$
g_k^- = g_{-k}, k \in \mathbb{Z}.
$$

Hence, if $G(z)$ denotes the z-transform of the filter $\mathbf{g}$, then the z-transform of the time reverse $\mathbf{g}^-$ of $\mathbf{g}$ is given by $G\left(\frac{1}{z}\right)$.

Definition 2 $\boxed{\textbf{Perfect-reconstruction filter banks}}$ *A pair $\{\widetilde{\boldsymbol{\ell}}, \widetilde{\mathbf{h}}\}$ of lowpass and highpass filters is said to have a PR dual pair $\{\boldsymbol{\ell}, \mathbf{h}\}$, if these two pairs provide a* "**perfect-reconstruction**" (PR) *filter bank, in the sense as described by the data-flow diagram in Fig. 8.8.*
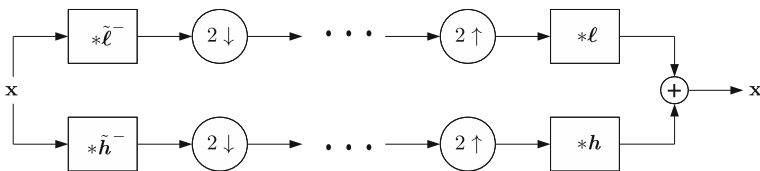
Fig. 8.8. Decomposition and reconstruction algorithms with PR filter bank

Observe that the PR dual of the Haar filter pair $\{\boldsymbol{\ell}, \mathbf{h}\}$ given in Example 4 is the pair $\{\frac{1}{2}\boldsymbol{\ell}, \frac{1}{2}\mathbf{h}\}$. Next we discuss the condition under which two pairs of digital filters constitute a PR filter bank.

Theorem 1   *Let $L(z), H(z), \widetilde{L}(z), \widetilde{H}(z)$ denote the z-transforms of $\boldsymbol{\ell}, \mathbf{h}, \widetilde{\boldsymbol{\ell}}, \widetilde{\mathbf{h}}$, respectively. Then the two filter pairs $\{\widetilde{\boldsymbol{\ell}}, \widetilde{\mathbf{h}}\}$ and $\{\boldsymbol{\ell}, \mathbf{h}\}$ constitute a PR filter bank, if and only if*

$$\left(L(z)\widetilde{L}\left(\frac{1}{z}\right) + H(z)\widetilde{H}\left(\frac{1}{z}\right)\right) X(z) + \left(L(z)\widetilde{L}\left(-\frac{1}{z}\right) + H(z)\widetilde{H}\left(-\frac{1}{z}\right)\right) X(-z)$$
$$= 2X(z) \tag{8.4.3}$$

for any bi-infinite sequence with z-transform $X(z)$.

Proof   Let us first consider the z-transform of $\mathbf{y} = UD\mathbf{w}$, obtained by downsampling, followed by upsampling, of any sequence $\mathbf{w} = \{w_j\}$. Then from the definitions of D and U, we have

$$\mathbf{y} = UD\mathbf{w} = (\dots, 0, w_{-2}, 0, w_0, 0, w_2, \dots);$$

that is, $y_{2j} = w_{2j}$, $y_{2j-1} = 0$, $j \in \mathbb{Z}$. Thus, the z-transform of $\mathbf{y}$ is given by

$$Y(z) = \sum_k y_k z^k = \sum_j y_{2j} z^{2j} = \sum_j w_{2j} z^{2j}.$$

On the other hand, since

$$\sum_j w_{2j} z^{2j} = \sum_k \frac{1 + (-1)^k}{2} w_k z^k$$
$$= \frac{1}{2}\left(\sum_k w_k z^k + \sum_k (-1)^k w_k z^k\right) = \frac{1}{2}(W(z) + W(-z)),$$

it follows that

$$Y(z) = \frac{1}{2}(W(z) + W(-z)).$$

In light of the above derivation, we may conclude that the z-transforms of $UD(\mathbf{x} * \widetilde{\boldsymbol{\ell}}^-)$ and $UD(\mathbf{x} * \widetilde{\mathbf{h}}^-)$ are given by

$$\frac{1}{2}\left(\widetilde{L}\left(\frac{1}{z}\right) X(z) + \widetilde{L}\left(-\frac{1}{z}\right) X(-z)\right), \quad \frac{1}{2}\left(\widetilde{H}\left(\frac{1}{z}\right) X(z) + \widetilde{H}\left(-\frac{1}{z}\right) X(-z)\right).$$

Therefore, the two pairs $\{\boldsymbol{\ell}, \mathbf{h}\}$ and $\{\widetilde{\boldsymbol{\ell}}, \widetilde{\mathbf{h}}\}$ constitute a PR filter bank $\{\boldsymbol{\ell}, \mathbf{h}\}, \{\widetilde{\boldsymbol{\ell}}, \widetilde{\mathbf{h}}\}$, if and only if

$$\frac{1}{2}\left(\widetilde{L}\left(\frac{1}{z}\right)X(z)+\widetilde{L}\left(-\frac{1}{z}\right)X(-z)\right)L(z)+\frac{1}{2}\left(\widetilde{H}\left(\frac{1}{z}\right)X(z)+\widetilde{H}\left(-\frac{1}{z}\right)X(-z)\right)H(z)$$
$$=X(z),$$

which is (8.4.3). ∎

We remark that from the symmetric formulation in (8.4.3), we may also conclude that if $\{\boldsymbol{\ell}, \mathbf{h}\}$ is a PR dual of $\{\widetilde{\boldsymbol{\ell}}, \widetilde{\mathbf{h}}\}$, then $\{\widetilde{\boldsymbol{\ell}}, \widetilde{\mathbf{h}}\}$ is a PR dual of $\{\boldsymbol{\ell}, \mathbf{h}\}$.

In order to apply (8.4.3) in Theorem 1 to verify if two filter pairs constitute a PR filter bank, it is necessary to show that (8.4.3) holds for all $X(z)$. This is not an easy (if feasible) task, since both $X(z)$ and $X(-z)$ appear in (8.4.3). To eliminate one of them, say $X(-z)$, in (8.4.3), we may additionally assume that the z-transforms of the two filter pairs $\{\boldsymbol{\ell}, \mathbf{h}\}$ and $\{\widetilde{\boldsymbol{\ell}}, \widetilde{\mathbf{h}}\}$ satisfy the identity:

$$L(z)\widetilde{L}\left(-\frac{1}{z}\right)+H(z)\widetilde{H}\left(-\frac{1}{z}\right)=0,$$

so that (8.4.3) reduces to the identity

$$L(z)\widetilde{L}\left(\frac{1}{z}\right)+H(z)\widetilde{H}\left(\frac{1}{z}\right)=2.$$

The above two identities can also be formulated by replacing z with $-z$, yielding the following four identities:

$$L(z)\widetilde{L}\left(\frac{1}{z}\right)+H(z)\widetilde{H}\left(\frac{1}{z}\right)=2, \tag{8.4.4a}$$

$$L(-z)\widetilde{L}\left(\frac{1}{z}\right)+H(-z)\widetilde{H}\left(\frac{1}{z}\right)=0, \tag{8.4.4b}$$

$$L(-z)\widetilde{L}\left(-\frac{1}{z}\right)+H(-z)\widetilde{H}\left(-\frac{1}{z}\right)=2, \tag{8.4.4c}$$

$$L(z)\widetilde{L}\left(-\frac{1}{z}\right)+H(z)\widetilde{H}\left(-\frac{1}{z}\right)=0. \tag{8.4.4d}$$

To re-formulate these 4 identities in a more compact form, we introduce the notion of modulation matrices, $M_{L,H}(z)$ and $M_{\widetilde{L},\widetilde{H}}(z)$, corresponding to the two pairs $\{L(z), H(z)\}$ and $\{\widetilde{L}(z), \widetilde{H}(z)\}$ of Laurent polynomials, defined by

$$M_{L,H}(z)=\begin{bmatrix}L(z)&L(-z)\\H(z)&H(-z)\end{bmatrix};\ M_{\widetilde{L},\widetilde{H}}(z)=\begin{bmatrix}\widetilde{L}(z)&\widetilde{L}(-z)\\\widetilde{H}(z)&\widetilde{H}(-z)\end{bmatrix}.$$

Then (8.4.4) (that is, the totality of (8.4.4a)–(8.4.4d)) is equivalent to the matrix identity:

$$\left(M_{L,H}(z)\right)^T \, M_{\widetilde{L},\widetilde{H}}\left(\frac{1}{z}\right) = 2I_2, \ z \in \mathbb{C}\backslash\{0\} \tag{8.4.5}$$

(see Exercise 7). Since the totality of (8.4.4a)–(8.4.4d) implies the validity of (8.4.3) for all $X(z)$, the following result is an immediate consequence of Theorem 1.

Theorem 2 | **Perfect-reconstruction filter banks** | *Suppose that the modulation matrices $M_{L,H}(z)$ and $M_{\widetilde{L},\widetilde{H}}(z)$ corresponding to the two z-transform pairs $\{L(z), H(z)\}$ and $\{\widetilde{L}(z), \widetilde{H}(z)\}$ of the filter pairs $\{\boldsymbol{\ell}, \mathbf{h}\}$ and $\{\widetilde{\boldsymbol{\ell}}, \widetilde{\mathbf{h}}\}$, respectively, satisfy the matrix identity (8.4.5). Then the filter bank $\{\widetilde{\boldsymbol{\ell}}, \widetilde{\mathbf{h}}, \boldsymbol{\ell}, \mathbf{h}\}$ constitutes a PR filter bank.*

Remark 1 | **Biorthogonal filters are PR filter banks** | Let $p = \{p_k\}, q = \{q_k\},$ $\widetilde{p} = \{\widetilde{p}_k\}, \widetilde{q} = \{\widetilde{q}_k\}$ be FIR filters, and set

$$\ell_k = p_k, \ h_k = q_k, \ \widetilde{\ell}_k = \frac{1}{2}\widetilde{p}_k, \ \widetilde{h}_k = \frac{1}{2}\widetilde{q}_k.$$

Then the decomposition and reconstruction algorithms (8.3.4)–(8.3.5) on p.406 can be formulated as

$$\mathbf{c} = (\widetilde{\boldsymbol{\ell}}^- * \mathbf{x}) \downarrow 2, \quad \mathbf{d} = (\widetilde{\mathbf{h}}^- * \mathbf{x}) \downarrow 2, \tag{8.4.6}$$

$$\widetilde{\mathbf{x}} = (\mathbf{c} \uparrow 2) * \boldsymbol{\ell} + (\mathbf{d} \uparrow 2) * \mathbf{h}, \tag{8.4.7}$$

where $\downarrow 2$ and $\uparrow 2$ denote the downsampling and upsampling operators. Therefore, if $p, q, \widetilde{p}, \widetilde{q}$ are biorthogonal, then the filter bank $\{\widetilde{\boldsymbol{\ell}}, \widetilde{\mathbf{h}}, \boldsymbol{\ell}, \mathbf{h}\}$ constitutes a PR filter bank. For this reason, the biorthogonal filters $p, q, \widetilde{p}, \widetilde{q}$ are also called PR filters. Now, let $P(z), Q(z), \widetilde{P}(z)$ and $\widetilde{Q}(z)$ be the two-scale symbols of $p, q, \widetilde{p}$ and $\widetilde{q}$, respectively; and let $L(z), H(z), \widetilde{L}(z)$ and $\widetilde{H}(z)$ be the z-transforms of $\ell, h, \widetilde{\ell}$ and $\widetilde{h}$, respectively. Then we have

$$P(z) = \frac{1}{2}L(z), \ Q(z) = \frac{1}{2}H(z),$$
$$\widetilde{P}(z) = \widetilde{L}(z), \ \widetilde{Q}(z) = \widetilde{H}(z).$$

Thus, the condition (8.3.11) on p.409 for the biorthogonality of $p, q, \widetilde{p}, \widetilde{q}$ is equivalent to (8.4.5). Hence, Theorem 2 can be derived directly from Remark 2 on p.410 in the previous section. ∎

Next, let us apply the above result to determine the existence of an FIR dual filter pair $\{\widetilde{\boldsymbol{\ell}}, \widetilde{\mathbf{h}}\}$ for a given FIR filter pair $\{\boldsymbol{\ell}, \mathbf{h}\}$. This problem is equivalent to studying the existence of Laurent polynomials $\widetilde{L}(z)$ and $\widetilde{H}(z)$ with modulation matrix $M_{\widetilde{L},\widetilde{H}}(z)$ that satisfies the matrix identity (8.4.5). To this end, we call a Laurent polynomial $w(z)$ a monomial, if $w(z) = cz^{-m}$ for some $m \in \mathbb{Z}$ and some constant $c \neq 0$.

Theorem 3 | **Existence of FIR PR dual** | *Let $\{\boldsymbol{\ell}, \mathbf{h}\}$ be an FIR filter pair with z-transform pair $\{L(z), H(z)\}$ and modulation matrix $M_{L,H}(z)$. Then there exists*

an FIR dual filter pair $\{\widetilde{\boldsymbol{\ell}}, \widetilde{\mathbf{h}}\}$ corresponding to $\{\boldsymbol{\ell}, \mathbf{h}\}$, provided that $\det\big(M_{L,H}(z)\big)$ *is a monomial.*

Proof   From the assumption on the modulation matrix $M_{L,H}(z)$, we may write

$$\det\big(M_{L,H}(z)\big) = cz^m$$

for some integer m and constant $c \neq 0$. First observe that by interchanging its two columns, the matrix $M_{L,H}(z)$ becomes $M_{L,H}(-z)$, so that $\det\big(M_{L,H}(-z)\big) = -cz^m$. On the other hand, by replacing z with $-z$ in the matrix, we also have $\det\big(M_{L,H}(-z)\big) = c(-z)^m$. Hence, m is an odd integer. Let us now introduce the two Laurent polynomials

$$\widetilde{L}(z) = \frac{2}{c}z^m H\left(-\frac{1}{z}\right), \quad \widetilde{H}(z) = -\frac{2}{c}z^m L\left(-\frac{1}{z}\right). \tag{8.4.8}$$

From the fact that m is an odd integer, it is easy to verify that the modulation matrix $M_{\widetilde{L},\widetilde{H}}(z)$ satisfies (8.4.5) (see Exercise 9). Thus, it follows from Theorem 2 that the filter bank $\{\widetilde{\boldsymbol{\ell}}, \widetilde{\mathbf{h}}, \boldsymbol{\ell}, \mathbf{h}\}$ constitutes a PR filter bank; that is, $\{\widetilde{\boldsymbol{\ell}}, \widetilde{\mathbf{h}}\}$ is a dual filter pair of $\{\boldsymbol{\ell}, \mathbf{h}\}$.   ∎

Example 5   Since the z-transforms $L(z), H(z)$ of the Haar filter pair $\{\boldsymbol{\ell}, \mathbf{h}\}$ in Example 4 are given by

$$L(z) = 1 + z, \quad H(z) = 1 - z,$$

it is easy to compute the determinant of the modulation matrix, namely:

$$\begin{aligned}
\det\big(M_{L,H}(z)\big) &= \det\left(\begin{bmatrix} L(z) & L(-z) \\ H(z) & H(-z) \end{bmatrix}\right) \\
&= (1+z)(1+z) - (1-z)(1-z) = 4z,
\end{aligned}$$

which is indeed a monomial of odd degree. Therefore, it follows from (8.4.8), with $m = 1, c = 4$, in the proof of Theorem 3, that

$$\begin{aligned}
\widetilde{L}(z) &= \frac{2}{4}z H\left(-\frac{1}{z}\right) = \frac{1}{2}z \cdot \left(1 - \frac{1}{-z}\right) = \frac{1}{2}(z+1), \\
\widetilde{H}(z) &= -\frac{2}{4}z L\left(-\frac{1}{z}\right) = -\frac{z}{2} \cdot \left(1 + \frac{1}{-z}\right) = \frac{1}{2}(-z+1).
\end{aligned}$$

That is, the PR dual of $\{\boldsymbol{\ell}, \mathbf{h}\}$ is the FIR filter pair $\{\frac{1}{2}\boldsymbol{\ell}, \frac{1}{2}\mathbf{h}\}$, which is precisely the time-reverse of $\{\mathbf{s}, \mathbf{r}\}$, as discussed in Example 2.   ∎

Let us now return to (8.4.5) and observe that the matrix $\frac{1}{2}\big(M_{L,H}(z)\big)^T$ is the left-inverse of $M_{\widetilde{L},\widetilde{H}}(\frac{1}{z})$. Since the right-inverse is the same as the left-inverse for

non-singular square matrices, we have

$$M_{\widetilde{L},\widetilde{H}}\left(\frac{1}{z}\right)\frac{1}{2}\left(M_{L,H}(z)\right)^{T} = I_2,$$

or equivalently,

$$M_{\widetilde{L},\widetilde{H}}(z)\left(M_{L,H}\left(\frac{1}{z}\right)\right)^{T} = 2I_2, \ z \in \mathbb{C}\setminus\{0\}.$$

Hence, the filter pair $\{\widetilde{\boldsymbol{\ell}}, \widetilde{\mathbf{h}}\}$ is dual to the filter pair $\{\boldsymbol{\ell}, \mathbf{h}\}$, if and only if

$$\begin{bmatrix} \widetilde{L}(z) & \widetilde{L}(-z) \\ \widetilde{H}(z) & \widetilde{H}(-z) \end{bmatrix} \cdot \begin{bmatrix} L\left(\frac{1}{z}\right) & H\left(\frac{1}{z}\right) \\ L\left(-\frac{1}{z}\right) & H\left(-\frac{1}{z}\right) \end{bmatrix} = 2I_2,$$

which is equivalent to a set of four identities:

$$\widetilde{L}(z)L\left(\frac{1}{z}\right) + \widetilde{L}(-z)L\left(-\frac{1}{z}\right) = 2, \tag{8.4.9a}$$

$$\widetilde{L}(z)H\left(\frac{1}{z}\right) + \widetilde{L}(-z)H\left(-\frac{1}{z}\right) = 0, \tag{8.4.9b}$$

$$\widetilde{H}(z)L\left(\frac{1}{z}\right) + \widetilde{H}(-z)L\left(-\frac{1}{z}\right) = 0, \tag{8.4.9c}$$

$$\widetilde{H}(z)H\left(\frac{1}{z}\right) + \widetilde{H}(-z)H\left(-\frac{1}{z}\right) = 2. \tag{8.4.9d}$$

Remark 2 $\boxed{\textbf{Lowpass/highpass dual filters and filter orthogonality}}$ Among the above four identities, (8.4.9a) describes the "duality" of the lowpass filters: $\boldsymbol{\ell}$ and $\widetilde{\boldsymbol{\ell}}$, while (8.4.9d) describes the "duality" of the highpass filters: $\mathbf{h}$ and $\widetilde{\mathbf{h}}$. On the other hand, (8.4.9b) and (8.4.9c) describe, respectively, the "orthogonality" between $\widetilde{\boldsymbol{\ell}}$ and $\mathbf{h}$ and the "orthogonality" between $\boldsymbol{\ell}$ and $\widetilde{\mathbf{h}}$. These four relations, with two on duality and two on orthogonality, are very useful for the design of PR filter banks (see Exercise 7). ∎

Example 6 Let

$$L(z) = 1 + z, \ H(z) = 1 - z, \ \widetilde{L}(z) = \frac{1}{2}(1 + z), \ \widetilde{H}(z) = \frac{1}{2}(1 - z)$$

be the z-transforms of the Haar filter bank $\{\boldsymbol{\ell}, \mathbf{h}\,\widetilde{\boldsymbol{\ell}}, \widetilde{\mathbf{h}}\}$ considered in Example 5. In this example we verify that these Laurent polynomials satisfy the identities (8.4.9a)–(8.4.9d), so that $\{\boldsymbol{\ell}, \widetilde{\boldsymbol{\ell}}\}$ is a lowpass dual filter pair, $\{\mathbf{h}, \widetilde{\mathbf{h}}\}$ is a highpass

dual filter pair, while $\widetilde{\ell}$ is "orthogonal" to $\mathbf{h}$, and ℓ is "orthogonal" to $\widetilde{\mathbf{h}}$. Indeed, first observe that

$$\widetilde{L}(z)L\left(\frac{1}{z}\right) = \frac{1}{2}(1+z)\left(1+\frac{1}{z}\right) = \frac{1}{2}\left(z+2+\frac{1}{z}\right),$$

so that $\widetilde{L}(-z)L\left(-\frac{1}{z}\right) = \frac{1}{2}\left(-z+2-\frac{1}{z}\right)$. Hence, we have

$$\widetilde{L}(z)L\left(\frac{1}{z}\right) + \widetilde{L}(-z)L\left(-\frac{1}{z}\right) = 2,$$

or $L(z)$ and $\widetilde{L}(z)$ satisfy (8.4.9a). Next, observe that

$$\widetilde{H}(z)H\left(\frac{1}{z}\right) = \frac{1}{2}(1-z)\left(1-\frac{1}{z}\right) = \frac{1}{2}\left(-z+2-\frac{1}{z}\right),$$

so that $\widetilde{H}(-z)H\left(-\frac{1}{z}\right) = \frac{1}{2}\left(z+2+\frac{1}{z}\right)$. Hence, we also have

$$\widetilde{H}(z)H\left(\frac{1}{z}\right) + \widetilde{H}(-z)H\left(-\frac{1}{z}\right) = 2,$$

or $H(z)$ and $\widetilde{H}(z)$ satisfy (8.4.9d). In addition, since

$$\widetilde{L}(z)H\left(\frac{1}{z}\right) = \frac{1}{2}(1+z)\left(1-\frac{1}{z}\right) = \frac{1}{2}\left(z-\frac{1}{z}\right),$$

so that $\widetilde{L}(-z)H\left(-\frac{1}{z}\right) = \frac{1}{2}\left(-z+\frac{1}{z}\right)$, we have

$$\widetilde{L}(z)H\left(\frac{1}{z}\right) + \widetilde{L}(-z)H\left(-\frac{1}{z}\right) = 0,$$

which implies that (8.4.9b) holds. Finally, since

$$\widetilde{H}(z)L\left(\frac{1}{z}\right) = \frac{1}{2}(1-z)\left(1+\frac{1}{z}\right) = \frac{1}{2}\left(-z+\frac{1}{z}\right),$$

so that $\widetilde{H}(-z)L\left(-\frac{1}{z}\right) = \frac{1}{2}\left(z-\frac{1}{z}\right)$, we also have

$$\widetilde{H}(z)L\left(\frac{1}{z}\right) + \widetilde{H}(-z)L\left(-\frac{1}{z}\right) = 0,$$

which implies that (8.4.9c) holds as well.   ∎

Exercises

Exercise 1 Compute the discrete convolution $\mathbf{x} * \mathbf{y}$ for each of the following pairs of bi-infinite sequences $\mathbf{x} = \{x_j\}$, $\mathbf{y} = \{y_j\}$.

(a) $x_j = \begin{cases} 2, & \text{for } j = 0, \\ -1, & \text{for } j = 1, \\ 0, & \text{otherwise;} \end{cases}$ $y_j = \begin{cases} -1, & \text{for } j = -1, \\ 2, & \text{for } j = 0, \\ -1, & \text{for } j = 1, \\ 0, & \text{otherwise.} \end{cases}$

(b) $x_j = \begin{cases} 1, & \text{for } j = 0, \\ 1/2, & \text{for } j = -1 \text{ or } 1, \\ 0, & \text{otherwise;} \end{cases}$ $y_j = \begin{cases} 1, & \text{for } j = 0 \text{ or } 1, \\ 0, & \text{otherwise.} \end{cases}$

(c) $\{x_j\}$ as given in (b), and $y_j = x_j$.

Exercise 2 In each of (a), (b), and (c) in Exercise 1, compute the z-transforms $X(z)$, $Y(z)$ and $W(z)$, where $\mathbf{w} = \mathbf{x} * \mathbf{y}$.

Exercise 3 Among all the sequences: $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{w} = \mathbf{x} * \mathbf{y}$ in each of (a), (b), and (c) in Exercise 1, identify those that are lowpass and those that are highpass filters.

Exercise 4 Let $\boldsymbol{\ell} = \{\ell_j\}$ and $\mathbf{h} = \{h_j\}$ be lowpass and highpass filters, respectively, as given in (a) and (b) below. For any bi-infinite sequence $\mathbf{x} = \{x_j\}$, write out the lowpass and highpass outputs $\mathbf{x}^L$ and $\mathbf{x}^H$ explicitly.

(a) $\ell_j = \begin{cases} 1/2, & \text{for } j = -1 \text{ or } 1, \\ 1, & \text{for } j = 0, \\ 0, & \text{otherwise;} \end{cases}$

$h_j = \begin{cases} -1/2, & \text{for } j = -1 \text{ or } 1, \\ 1, & \text{for } j = 0, \\ 0, & \text{otherwise.} \end{cases}$

(b) $\boldsymbol{\ell}$ is the same as in (a), and $h_j = \begin{cases} -1/2, & \text{for } j = -1, \\ 1/2, & \text{for } j = 0, \\ 0, & \text{otherwise.} \end{cases}$

Exercise 5 As a continuation of Exercise 4, give an explicit formulation of $\mathbf{x}^{LD}$ and $\mathbf{x}^{HD}$ for each of (a) and (b).

Exercise 6 As a continuation of Exercise 5, give an explicit formulation of $U\mathbf{x}^{LD}$ and $U\mathbf{x}^{HD}$ for each of (a) and (b).

Exercise 7 Show that (8.4.4) [that is, the totality of (8.4.4a)–(8.4.4d)] is equivalent to the matrix identity (8.4.5). Also, show that for any non-singular square matrix A, if a square matrix B satisfies $AB = I$, then $BA = I$. Finally, apply this fact to show that the totality of the identities (8.4.4a)–(8.4.4d) is equivalent to the totality of the identities (8.4.9a)–(8.4.9d).

Exercise 8 Write out the details in the derivation of (8.4.4a) and (8.4.4b).

Exercise 9 Let $M_{L,H}(z)$ be the modulation matrix of $L(z)$, $H(z)$. Suppose

$$\det\left(M_{L,H}(z)\right) = cz^m$$

for some odd integer m and constant $c \neq 0$. Show that $\widetilde{L}(z)$, $\widetilde{H}(z)$, defined by (8.4.8), satisfy (8.4.5).

Exercise 10 Let $\boldsymbol{\ell} = \{\ell_j\}$, $\widetilde{\boldsymbol{\ell}} = \{\widetilde{\ell}_j\}$, $\mathbf{h} = \{h_j\}$ and $\widetilde{\mathbf{h}} = \{\widetilde{h}_j\}$ with z-transforms $L(z)$, $\widetilde{L}(z)$, $H(z)$, and $\widetilde{H}(z)$, respectively. Formulate the duality condition (8.4.4a) and (8.4.4b) for the two pairs $\{\boldsymbol{\ell}, \mathbf{h}\}$ and $\{\widetilde{\boldsymbol{\ell}}, \widetilde{\mathbf{h}}\}$ in the time-domain; that is, formulate the condition in terms of ℓ_j, $\widetilde{\ell}_j$, h_j, $\widetilde{h}_j$ instead of $L(z)$, $\widetilde{L}(z)$, $H(z)$, $\widetilde{H}(z)$.

Exercise 11 Use the same notations as in Exercise 10 to formulate the following conditions in the time-domain: (a) the duality condition (8.4.9a) of $\boldsymbol{\ell}$ and $\widetilde{\boldsymbol{\ell}}$; (b) the duality condition (8.4.9d) of $\mathbf{h}$ and $\widetilde{\mathbf{h}}$; (c) the orthogonality condition (8.4.9b) of $\widetilde{\boldsymbol{\ell}}$ and $\mathbf{h}$, and (d) the orthogonality condition (8.4.9c) of $\boldsymbol{\ell}$ and $\widetilde{\mathbf{h}}$.

Chapter 9
Compactly Supported Wavelets

The objective of this chapter is three-fold: firstly to discuss the theory and methods for the construction of compactly supported wavelets, including both orthogonal and biorthogonal wavelets as well as their corresponding scaling functions; secondly to study the stability and smoothness properties of these basis functions; and thirdly to investigate the details in the derivation of fast computational algorithms, the lifting schemes, for the implementation of wavelet decompositions and reconstructions. To accomplish this goal, our preliminary task is to introduce the concept of the transition operator T_p, along with its representation matrix $\mathcal{T}_p$, associated with a given finite sequence p, as well as to study the properties of the eigenvalues and eigenvectors of the matrix $\mathcal{T}_p$. This will be discussed in Sect. 9.1, which also includes the Fundamental Lemma of transition operators that will be applied in the next chapter to prove the existence of the refinable function $\phi \in L_2(\mathbb{R})$ with p as its refinement mask, as well as to characterize the stability of ϕ, again in terms of the transition operator.

To facilitate our discussions, the Gramian function $G_{\widetilde{\phi},\phi}$, defined in terms of the Fourier transform of a pair of refinable functions, $\widetilde{\phi}$ and ϕ, by

$$G_{\widetilde{\phi},\phi}(\omega) = \sum_{\ell \in \mathbb{Z}} \widehat{\widetilde{\phi}}(\omega + 2\pi\ell)\overline{\widehat{\phi}(\omega + 2\pi\ell)},$$

is introduced in Sect. 9.2, where it is shown that

$$G_{\widetilde{\phi},\phi}(\omega) = \sum_{k} \int_{-\infty}^{\infty} \widetilde{\phi}(x)\overline{\phi(x-k)}dx \ e^{-ik\omega}.$$

Hence, for refinement masks with only finitely many non-zero terms, since the corresponding refinable functions are compactly supported, the Gramian functions are trigonometric polynomials. When only one refinable function ϕ, corresponding to

some refinement mask, is considered, it is also shown in this section that ϕ is stable, if and only if the Gramian function $G_\phi = G_{\phi,\phi}$ is uniformly bounded both from above and from below by some positive constants. As a consequence, the refinable function ϕ is a scaling function, as introduced in Definition 1, in Sect. 8.2 of Chap. 8. In particular, ϕ is an orthonormal (i.e. orthogonal and with $L_2(\mathbb{R})$ norm $= 1$) scaling function, if and only if G_ϕ is the constant function 1 (almost everywhere, if ϕ is not compactly supported).

In Sect. 9.3, the MRA (multiresolution approximation) architecture generated by a compactly supported orthonormal scaling function ϕ, as introduced in Definition 1 in Sect. 8.2 of Chap. 8, is extended to an orthogonal multiresolution analysis (also denoted by MRA) architecture for the formulation of compactly supported orthogonal wavelets in terms of ϕ. When the trigonometric polynomial $p(\omega)$ (i.e. the two-scale symbol $P(e^{-i\omega})$) of some finite sequence p satisfies the quadrature mirror filter (QMF) identity, it is shown that the trigonometric polynomial

$$q(\omega) = -p(\pi - \omega)e^{-i(2L-1)\omega}$$

is the two-scale symbol of a compactly supported wavelet corresponding to the scaling function ϕ (for an arbitrary integer L), provided that the transition operator T_p satisfies Condition E, in that the integer 1 is a simple eigenvalue of the representation matrix $\mathcal{T}_p$ of T_p and any other eigenvalue λ of $\mathcal{T}_p$ satisfies $|\lambda| < 1$. The notions of sum rules and vanishing moments are also introduced in Sect. 9.3, and it is proved that the sum-rule property of a given order implies the same order of vanishing moments for orthogonal wavelets. This section ends with the derivation of the method, along with formulas, for constructing compactly supported orthogonal wavelets, and with examples of the wavelet filters D_{2n} for $n = 1, \ldots, 4$.

Since compactly supported orthogonal wavelets, with the exception of the Haar wavelet, are not symmetric or anti-symmetric, and hence their corresponding wavelet filters are not linear-phase filters, the next section, Sect. 9.4, is devoted to the investigation of biorthogonal wavelets by using two different compactly supported scaling functions $\widetilde{\phi}$ and ϕ that are symmetric. For this study, "orthogonal basis" is replaced by the more general "Riesz basis". The weaker condition for "Bessel sequence" is introduced, by dropping the lower-bound requirement for Riesz bases. It is proved in Sect. 9.4, however, that any pair of biorthogonal systems must satisfy the lower-bound condition, if they are Bessel sequences. Hence, to derive biorthogonal wavelets, it is sufficient to verify the biorthogonality property and the upper-bound condition. The good news in this regard is that, under a very mild smoothness condition, any compactly supported wavelet ψ with vanishing moment of order (at least) one (i.e. the integral of ψ over its support is zero) already generates a Bessel system $\{\psi_{j,k} : j, k \in \mathbb{Z}\}$, where

$$\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k).$$

To derive the biorthogonality property of a pair $\{\widetilde{\phi}, \phi\}$ of refinable functions corresponding to the pair $\{\widetilde{p}, p\}$ of compactly supported refinement masks, it is again

necessary and sufficient to assure that the Gramian function $G_{\widetilde{\phi},\phi}$ is the constant function 1. This is accomplished by extending the QMF identity to the dual QMF:

$$\widetilde{p}(\omega)\overline{p(\omega)} + \widetilde{p}(\omega + \pi)\overline{p(\omega + \pi)} = 1,$$

where $\widetilde{p}(\omega)$ and $p(\omega)$ are the two-scale symbols of $\widetilde{p}$ and p, respectively. Furthermore, the corresponding biorthogonal wavelets $\widetilde{\psi}$ and ψ can be formulated via the two-scale symbols $\widetilde{q}(\omega)$ and $q(\omega)$ by applying the three identities:

$$\widetilde{q}(\omega)\overline{q(\omega)} + \widetilde{q}(\omega + \pi)\overline{q(\omega + \pi)} = 1,$$
$$\widetilde{q}(\omega)\overline{p(\omega)} + \widetilde{q}(\omega + \pi)\overline{p(\omega + \pi)} = 0,$$
$$\widetilde{p}(\omega)\overline{q(\omega)} + \widetilde{p}(\omega + \pi)\overline{q(\omega + \pi)} = 0.$$

This section ends with a description of the method for constructing compactly supported symmetric (or anti-symmetric) biorthogonal wavelets, and with examples of the 5/3-tap and 9/7-tap filters that have been adopted by the JPEG-2000 digital image compression standard, for lossless and lossy compressions, respectively.

Perhaps the most significant impact of the wavelet transform (WT) to the application of local time-scale and time-frequency analyses is that the (wavelet) decomposition algorithm eliminates the need for computing the WT integral in (8.1.5) to obtain the DWT values at the dyadic points, and that the (wavelet) reconstruction algorithm can be applied to recover the original input data. Therefore, fast computational schemes for implementation of the (wavelet) decomposition and reconstruction algorithms are of great interest. In Sect. 8.3 of Chap. 8, we have demonstrated the design of such implementation with two simple examples: the Haar and 5/3-tap wavelet filters, by introducing the notion of lifting steps. In the last section, Sect. 9.5, of this chapter, we will derive the general lifting schemes and establish the corresponding relevant results. Recall from Sect. 8.3 of Chap. 8 that the modulation matrices play an essential role in formulating the wavelet decomposition/reconstruction algorithms. Indeed, the wavelet filter pairs, $\{\widetilde{p}, \widetilde{q}\}$ and $\{p, q\}$, constitute a PR filter bank, if and only if their corresponding modulation matrices $M_{\widetilde{p},\widetilde{q}}(\omega)$ and $M_{p,q}(\omega)$ satisfy the identity

$$\overline{M_{\widetilde{p},\widetilde{q}}(\omega)}M_{p,q}(\omega)^T = I_2, \quad \omega \in \mathbb{R}.$$

Here, the first column of $M_{p,q}(\omega)$ is given by $[p(\omega), q(\omega)]^T = [P(z), Q(z)]^T$, where $z = e^{-i\omega}$, while the second column is obtained from the first column by replacing z with $e^{-i\pi}z$; that is, by a phase shift of π. In general, when the wavelet scale (also called dilation) 2 is generalized to an arbitrary integer $d \geq 2$, then instead of one wavelet filter q, it is necessary to consider $d - 1$ wavelet (highpass) filters $q_1, \ldots, q_{d-1}$. In this case, the first column of the corresponding modulation matrix $M_{p,q_1,\ldots,q_{d-1}}(\omega)$ is given by $[p(\omega), q_1(\omega), \ldots, q_{d-1}(\omega)]^T = [P(z), Q_1(z), \ldots, Q_{d-1}(z)]^T$, and the $(k + 1)$th column is obtained from the first column by a phase shift of $2\pi k/d$ or replacing z by $e^{-i2\pi k/d}z$. Observe that all columns of $M_{p,q_1,\ldots,q_{d-1}}(\omega)$ are correlated, in that they are phase shifts of one another. This makes it very difficult, if even feasible

at all, to construct the filters directly from the matrix identity:

$$\overline{M_{\widetilde{p},\widetilde{q}_1,\ldots,\widetilde{q}_{d-1}}(\omega)} M_{p,q_1,\ldots,q_{d-1}}(\omega)^T = I_d.$$

Therefore, the method of polyphase decomposition must first be applied to untangle the correlations. In Sect. 9.5, we only consider dilation $d = 2$, but still call the bi-phase matrices by the commonly used term, polyphase matrices. The main result of this section is the factorization of the DWT decomposition/reconstruction algorithms into lifting steps. Examples, including the lifting schemes for implementation of the D_4, 5/3-tap and 9/7-tap DWT, will be discussed at the end of the section.

9.1 Transition Operators

The notion of the transition operator and its representation matrix, associated with a given refinement mask, is introduced in this first section. In particular, the properties of the eigenvalues and eigenvectors of the transition operator will be investigated, since they are instrumental to the characterization of the orthogonality, stability, and smoothness of the scaling function corresponding to the refinement mask, to be studied in the later sections.

In this chapter, we will only consider (refinement) sequences with compact support; that is, sequences with only finitely many non-zero terms. For simplicity of our discussion, we may and will assume, without loss of generality, that the support of the sequence $p = \{p_k\}$ under consideration is $[0, N]$ for some $N \in \mathbb{N}$; that is, $p_k = 0$ for all $k < 0$ or $k > N$. (As in the previous chapter, when a sequence $p = \{p_k\}$ is considered as a refinement sequence, we do not bold-face p.) For a sequence p with support on $[0, N]$, we will call $[0, N]$ the support of $p = \{p_k\}$, if both p_0 and p_N are different from 0, and will use the notation $p = \{p_k\}_{k=0}^N$ for p, if needed for clarity. Thus, the two-scale symbol $P(z)$ of this sequence p is given by

$$P(z) = \frac{1}{2} \sum_{k=0}^N p_k z^k, \quad z \in \mathbb{C}\setminus\{0\}.$$

In this and the next chapters, we will also use the notation $p(\omega)$ to replace $P(e^{-i\omega})$ for convenience, namely:

$$p(\omega) = \frac{1}{2} \sum_{k=0}^N p_k e^{-ik\omega}, \tag{9.1.1}$$

so that $p(\omega)$ is a trigonometric polynomial. Suppose that there exists some function ϕ that satisfies the identity:

$$\phi(x) = \sum_k p_k \phi(2x - k), \tag{9.1.2}$$

then ϕ is a called a **refinable function** with refinement sequence (also called refinement mask) $p = \{p_k\}$. An equivalent formulation of the identity (9.1.2) is the following frequency-domain representation:

$$\widehat{\phi}(\omega) = p\left(\frac{\omega}{2}\right)\widehat{\phi}\left(\frac{\omega}{2}\right). \tag{9.1.3}$$

Throughout this and the next chapters, a refinement mask $p = \{p_k\}$ is always assumed to satisfy the condition:

$$\sum_k p_k = 2,$$

or equivalently, $P(1) = p(0) = 1$. In addition, as in the previous chapter, when $p = \{p_k\}$, $\widetilde{p} = \{\widetilde{p}_k\}$ and $q = \{q_k\}$, $\widetilde{q} = \{\widetilde{q}_k\}$ are considered as the refinement and wavelet sequences, we do not bold-face p, $\widetilde{p}$, q, $\widetilde{q}$ (see Remark 1 on p.396).

Remark 1 The definition of refinable functions in (9.1.2) is valid for all (finite or infinite) sequences $p = \{p_k\}$. Of course, if $p = \{p_k\}_{k=0}^{N}$ has compact support, then the summation in (9.1.2) runs from $k = 0$ to $k = N$. But in any case, ϕ is called a refinable function with refinement mask $p = \{p_k\}$. Furthermore, the definition of refinable functions is more general than that of scaling functions, introduced in Definition 1 in Sect. 8.2 of Chap. 8, in that refinable functions do not necessarily satisfy the stability condition, which is required for scaling functions. Hence, while the Fourier transform $\widehat{\phi}(\omega)$ of a scaling function must satisfy the condition $\widehat{\phi}(0) = c$ for a nonzero constant c, this is not necessarily the case for refinable functions in general. However, in the following, a refinable function ϕ associated with $p = \{p_k\}$ means the normalized solution of (9.1.2), given by (8.2.21) in Sect. 8.2, with the condition $\widehat{\phi}(0) = 1$. ∎

We are now ready to introduce the notion of the transition operator. For a refinement mask $p = \{p_k\}$ with compact support, the **transition operator** T_p associated with p is defined by

$$(T_p v)(\omega) = \left|p\left(\frac{\omega}{2}\right)\right|^2 v\left(\frac{\omega}{2}\right) + \left|p\left(\frac{\omega}{2} + \pi\right)\right|^2 v\left(\frac{\omega}{2} + \pi\right), \tag{9.1.4}$$

for all trigonometric polynomials $v(\omega)$. As in Sect. 6.1 of Chap. 6, let $\mathbb{V}_{2N+1}$ be the finite-dimensional linear space defined by

$$\mathbb{V}_{2N+1} = \left\{v(\omega): \ v(\omega) = \sum_{k=-N}^{N} v_k e^{-ik\omega}, \ v_k \in \mathbb{C}\right\}$$

(see (6.1.2) in Chap. 6). We have the following result.

Theorem 1 | **Representation matrix of transition operator** | *The transition ope-*

rator T_p associated with $p = \{p_k\}_{k=0}^{N}$ has the following properties:

(i) *the space $\mathbb{V}_{2N+1}$ is invariant under the transformation T_p; that is, $T_p v \in \mathbb{V}_{2N+1}$ for any $v \in \mathbb{V}_{2N+1}$;*

(ii) *with respect to the basis $\{e^{-ik\omega}\}_{k=-N,\dots,N}$ for $\mathbb{V}_{2N+1}$, the representation matrix of T_p, restricted to $\mathbb{V}_{2N+1}$, is given by*

$$\mathcal{T}_p = \left[\frac{1}{2} a_{2j-k} \right]_{-N \leq j,k \leq N},$$

where

$$a_j = \sum_{s \in \mathbb{Z}} p_s \overline{p}_{s-j} = \sum_{s=0}^{N} p_s \overline{p}_{s-j}. \tag{9.1.5}$$

More explicitly,

$$(T_p E)(\omega) = \left[T_p(e^{-ik\omega}) \right]_{k=-N,\dots,N} = E(\omega)\mathcal{T}_p, \tag{9.1.6}$$

where

$$E(\omega) = \left[e^{-ik\omega} \right]_{k=-N,\dots,N} = \left[e^{iN\omega}, \dots, e^{i\omega}, 1, e^{-i\omega}, \dots, e^{-iN\omega} \right].$$

Proof   First observe that since $p_k = 0$ for $k < 0$ or $k > N$, we have $a_j = 0$ if $j < -N$ or $j > N$. Thus, for any k with $0 \leq k \leq N$, $a_{2j-k} = 0$ if $j < -N$ or $j > N$. Next, we note that

$$|p(\omega)|^2 = \frac{1}{2} \sum_s p_s e^{-is\omega} \frac{1}{2} \sum_k \overline{p}_k e^{ik\omega} = \frac{1}{4} \sum_{s,k} p_s \overline{p}_k e^{-i(s-k)\omega}$$

$$= \frac{1}{4} \sum_s \sum_j p_s \overline{p}_{s-j} e^{-ij\omega} = \frac{1}{4} \sum_j \left(\sum_s p_s \overline{p}_{s-j} \right) e^{-ij\omega}$$

$$= \frac{1}{4} \sum_j a_j e^{-ij\omega}.$$

Thus, for $v(\omega) = e^{-ik\omega}$, where $-N \leq k \leq N$, we have

$$T_p(e^{-ik\omega}) = \frac{1}{4} \sum_s a_s e^{-is\omega/2} e^{-ik\omega/2} + \frac{1}{4} \sum_s a_s e^{-is(\omega/2+\pi)} e^{-ik(\omega/2+\pi)}$$

$$= \frac{1}{4} \sum_s a_s e^{-i(s+k)\omega/2} + \frac{1}{4} \sum_s a_s (-1)^{s+k} e^{-i(s+k)\omega/2}$$

$$= \frac{1}{4} \sum_{\ell} a_{\ell-k} \left(1 + (-1)^{\ell}\right) e^{-i\ell\omega/2} \quad \text{(with } \ell = s + k)$$

$$= \sum_{j} \frac{1}{2} a_{2j-k} e^{-ij\omega} \quad \text{(only even terms } \ell = 2j \text{ are kept)} \quad (9.1.7)$$

$$= \sum_{j=-N}^{N} \frac{1}{2} a_{2j-k} e^{-ij\omega} \quad \text{(since } a_{2j-k} = 0 \quad \text{if} \quad |j| > N)$$

$$= E(\omega) \left[\frac{1}{2} a_{2j-k} \right]_{j=N}^{-N}.$$

Therefore,

$$T_p \left[e^{iN\omega}, \ldots, e^{-ik\omega}, \ldots, e^{-iN\omega} \right]$$

$$= E(\omega) \left[\frac{1}{2} a_{2j-k} \right]_{-N \leq j \leq N, -N \leq k \leq N} = E(\omega) \mathcal{T}_p.$$

This completes the proof of (9.1.6). ∎

Remark 2 To compute each entry a_j of $\mathcal{T}_p$, although one may simply use its definition in (9.1.5), it is often easier to apply the relation

$$\sum_{j} a_j e^{-ij\omega} = 4|p(\omega)|^2 \tag{9.1.8}$$

and compute the Fourier coefficients (see the above proof of Theorem 1). ∎

Example 1 Let

$$\phi(x) = \min\{x, 2 - x\} \chi_{[0,2)}(x) = \begin{cases} x, & \text{for } 0 \leq x < 1, \\ 2 - x, & \text{for } 1 \leq x < 2, \\ 0, & \text{elsewhere,} \end{cases}$$

be the hat function. Recall from Example 2 in Sect. 8.2 of Chap. 8, on p.397, that ϕ is a refinable function with refinement mask:

$$p_0 = \frac{1}{2}, \ p_1 = 1, \ p_2 = \frac{1}{2},$$

and $p_k = 0$ for $k < 0$ or $k > 2$. Formulate the representation matrix $\mathcal{T}_p$ of the transition operator T_p (restricted to $\mathbb{V}_5$) associated with $p = \{p_k\}_{k=0}^2$.

Solution To compute a_j from (9.1.5), we have

$$a_{-2} = p_0 p_2 = \frac{1}{4}, \ a_{-1} = p_0 p_1 + p_1 p_2 = 1,$$

$$a_0 = p_0^2 + p_1^2 + p_2^2 = \frac{3}{2},$$

$$a_1 = p_1 p_0 + p_2 p_1 = 1, \quad a_2 = p_2 p_0 = \frac{1}{4},$$

and $a_k = 0$ for $|k| > 2$. Thus

$$
\mathcal{T}_p = \left[\frac{1}{2} a_{2j-k} \right]_{-2 \leq j,k \leq 2} = \frac{1}{2}
\begin{bmatrix}
a_{-2} & a_{-3} & a_{-4} & a_{-5} & a_{-6} \\
a_0 & a_{-1} & a_{-2} & a_{-3} & a_{-4} \\
a_2 & a_1 & a_0 & a_{-1} & a_{-2} \\
a_4 & a_3 & a_2 & a_1 & a_0 \\
a_6 & a_5 & a_4 & a_3 & a_2
\end{bmatrix}
$$

$$
= \frac{1}{8}
\begin{bmatrix}
1 & 0 & 0 & 0 & 0 \\
6 & 4 & 1 & 0 & 0 \\
1 & 4 & 6 & 4 & 1 \\
0 & 0 & 1 & 4 & 6 \\
0 & 0 & 0 & 0 & 1
\end{bmatrix}.
$$

∎

Remark 3 $\boxed{\textbf{Proper subspace} \mathbb{V}_{2N-1} \textbf{is also invariant under} T_p}$ Let T_p be the transition operator associated with $p = \{p_k\}_{k=0}^N$. For integers $k \in [1 - N, N - 1]$ and j with $|j| \geq N$, we have

$$|2j - k| \geq N + 1.$$

Thus $a_{2j-k} = 0$ for $k \in [1 - N, N - 1]$ and $|j| \geq N$. Therefore, by (9.1.7), we have, for an integer $k \in [1 - N, N - 1]$,

$$T_p(e^{-ik\omega}) = \sum_j \frac{1}{2} a_{2j-k} e^{-ij\omega} = \sum_{j=1-N}^{N-1} \frac{1}{2} a_{2j-k} e^{-ij\omega},$$

that is, $T_p(e^{-ik\omega}) \in \mathbb{V}_{2N-1}$. This means that $\mathbb{V}_{2N-1}$ is invariant under T_p, and the representation matrix of T_p, restricted to $\mathbb{V}_{2N-1}$, is given by

$$\left[\frac{1}{2} a_{2j-k} \right]_{1-N \leq j,k \leq N-1}.$$

∎

For a general refinement mask $p = \{p_k\}_{k=N_1}^{N_2}$, one can show, as above, that both $\mathbb{V}_{2(N_2-N_1)+1}$ and $\mathbb{V}_{2(N_2-N_1)-1}$ are invariant under T_p. In fact, for any $M \geq N_2 - N_1 - 1$, $\mathbb{V}_{2M+1}$ is invariant under T_p (just by considering $h = \{p_k\}_{k=N_1}^{M+N_1}$ with $p_k = 0$ for $k > N_2$ as the mask and observe $T_p = T_h$). In the next theorem, we will show that the nonzero eigenvalues of $T_p|_{\mathbb{V}_{2M+1}}$ (that is, the restriction of T_p to $\mathbb{V}_{2M+1}$) are the same as those of $T_p|_{\mathbb{V}_{2(N_2-N_1)+1}}$ for any $M > N_2 - N_1$.

For a trigonometric polynomial $v(\omega) = \sum_k v_k e^{-ik\omega}$, its **support** is defined by

$$\text{supp}(v) = [k_1, k_2],$$

provided that $v(\omega) = \sum_{k=k_1}^{k_2} v_k e^{-ik\omega}$ with $k_1 \leq k_2$ and $v_{k_1} \neq 0$, $v_{k_2} \neq 0$. A trigonometric polynomial $v(\omega)$ is called an eigenfunction of T_p associated with eigenvalue λ if

$$(T_p v)(\omega) = \lambda\, v(\omega).$$

Theorem 2 $\boxed{\textbf{Eigenfunctions of transition operator}}$ *Let T_p be the transition operator associated with $p = \{p_k\}_{k=N_1}^{N_2}$. Then an eigenfunction of T_p associated with a nonzero eigenvalue lies in $\mathbb{V}_{2(N_2-N_1)+1}$.*

Proof Suppose that the trigonometric polynomial $v(\omega) = \sum_k v_k e^{-ik\omega}$ is an eigenfunction of T_p associated with a nonzero eigenvalue λ. For an integer k, we have (by referring to the derivation of (9.1.7)),

$$T_p(e^{-ik\omega}) = \sum_j \frac{1}{2} a_{2j-k}\, e^{-ij\omega}.$$

Thus,

$$T_p v(\omega) = \sum_j \sum_k \frac{1}{2} a_{2j-k} v_k\, e^{-ij\omega}.$$

Observe that $a_{2j-k} = 0$ if $|2j - k| > N_2 - N_1$, or equivalently, a_{2j-k} is not zero only if

$$-\frac{N_2 - N_1}{2} + \frac{k}{2} \leq j \leq \frac{N_2 - N_1}{2} + \frac{k}{2}.$$

This means that

$$\text{supp}(T_p v) \subseteq \frac{1}{2}[N_1 - N_2, N_2 - N_1] + \frac{1}{2}\text{supp}(v).$$

Similarly, we have

$$\text{supp}(T_p^2 v) \subseteq \frac{1}{2}[N_1 - N_2, N_2 - N_1] + \frac{1}{2}\text{supp}(T_p v)$$

$$\subseteq \frac{1}{2}[N_1 - N_2, N_2 - N_1] + \frac{1}{2^2}[N_1 - N_2, N_2 - N_1] + \frac{1}{2^2}\text{supp}(v),$$

and, in general,

$$\text{supp}(T_p^n v)$$

$$\subseteq \frac{1}{2}[N_1 - N_2, N_2 - N_1] + \ldots + \frac{1}{2^n}[N_1 - N_2, N_2 - N_1] + \frac{1}{2^n}\text{supp}(v)$$

$$\subseteq [N_1 - N_2, N_2 - N_1] + \frac{1}{2^n}\text{supp}(v).$$

On the other hand, $T_p^n v = \lambda^n v$ with $\lambda \neq 0$ for $n \geq 0$. Thus $\text{supp}(T_p^n v) = \text{supp}(v)$. Therefore,

$$\text{supp}(v) = \text{supp}(T_p^n v) \subseteq [N_1 - N_2, N_2 - N_1] + \frac{1}{2^n} \text{supp}(v).$$

Letting $n \to \infty$ in the above equation, and using the fact that $\text{supp}(v)$ is bounded, we conclude that

$$\text{supp}(v) \subseteq [N_1 - N_2, N_2 - N_1].$$

This shows $v(\omega) \in \mathbb{V}_{2(N_2-N_1)+1}$, as desired.   ∎

Remark 4 In the next sections, the property of the eigenvalues (and in some cases, eigenvectors as well) of the transition operator will be used to characterize the existence, stability, orthogonality, and smoothness of scaling functions ϕ. From Theorem 2, we see that for a refinement mask $\{p_k\}_{k=N_1}^{N_2}$, we need only consider the transition operator T_p restricted to $\mathbb{V}_{2(N_2-N_1)+1}$, instead of $\mathbb{V}_{2M+1}$ for an $M > N_2 - N_1$. In the following, we simplify notations by using T_p for $T_p|_{\mathbb{V}_{2(N_2-N_1)+1}}$, which is the restriction of T_p to $\mathbb{V}_{2(N_2-N_1)+1}$. Hence the operator T_p is defined on the finite-dimensional space $\mathbb{V}_{2(N_2-N_1)+1}$, and the representation matrix of T_p is given by

$$\mathcal{T}_p = \left[\frac{1}{2} a_{2j-k} \right]_{N_1-N_2 \leq j, k \leq N_2-N_1}.$$

∎

Now let us return to the refinement mask supported on $[0, N]$; namely, $p = \{p_k\}_{k=0}^N$. The transition operator T_p is understood to be the linear operator defined on $\mathbb{V}_{2N+1}$.

For $v(\omega) = \sum_{j=-N}^N v_k e^{-ik\omega} \in \mathbb{V}_{2N+1}$, let vec (v) denote the column vector consisting of the coefficients v_k of $v(\omega)$, namely:

$$\text{vec}(v) = \left[v_k \right]_{k=N}^{-N} = \begin{bmatrix} v_{-N} \\ \vdots \\ v_k \\ \vdots \\ v_N \end{bmatrix}. \tag{9.1.9}$$

Then

$$v(\omega) = E(\omega)\text{vec}(v).$$

From (9.1.6), we have

$$T_p v(\omega) = T_p E(\omega)\text{vec}(v) = E(\omega)\mathcal{T}_p \text{vec}(v), \tag{9.1.10}$$

and thus

$$\text{vec}\left(T_p v\right) = \mathcal{T}_p \text{vec}(v). \tag{9.1.11}$$

Remark 5 | **Eigenfunctions of T_p and eigenvectors of $\mathcal{T}_p$** | If $v(\omega) \in \mathbb{V}_{2N+1}$ is an eigenfunction of T_p associated with an eigenvalue λ; that is, $T_p v(\omega) = \lambda v(\omega)$, then it follows from (9.1.11) that

$$\mathcal{T}_p \text{vec}(v) = \text{vec}\left(T_p v\right) = \text{vec}(\lambda v) = \lambda \text{vec}(v).$$

This means that λ is an eigenvalue of $\mathcal{T}_p$, with vec (v) being an associated eigenvector. Conversely, suppose $\mathbf{v}$ is a (right) eigenvector of $\mathcal{T}_p$ associated with an eigenvalue λ, i.e. $\mathcal{T}_p \mathbf{v} = \lambda \mathbf{v}$. Let $v(\omega) = E(\omega)\mathbf{v} \in \mathbb{V}_{2N+1}$. Then $\text{vec}(v) = \mathbf{v}$. Thus, by (9.1.10), we have

$$T_p v(\omega) = E(\omega)\mathcal{T}_p \mathbf{v} = \lambda E(\omega)\mathbf{v} = \lambda v(\omega).$$

That is, λ is an eigenvalue of T_p, with $v(\omega) = E(\omega)\mathbf{v} \in \mathbb{V}_{2N+1}$ being an associated eigenfunction. ∎

Example 2 As a continuation of Example 1, let us calculate the eigenvalues of $\mathcal{T}_p$ in Example 1. Observing that each of the first and the last rows of $\mathcal{T}_p$ has only one nonzero entry, we have

$$\det(\lambda I_5 - \mathcal{T}_p) = \left(\lambda - \frac{1}{8}\right)^2 \begin{vmatrix} \lambda - \frac{1}{2} & -\frac{1}{8} & 0 \\ -\frac{1}{2} & \lambda - \frac{3}{4} & -\frac{1}{2} \\ 0 & -\frac{1}{8} & \lambda - \frac{1}{2} \end{vmatrix}$$

$$= \left(\lambda - \frac{1}{8}\right)^2 (\lambda - 1)\left(\lambda - \frac{1}{2}\right)\left(\lambda - \frac{1}{4}\right),$$

where the details of the calculation for the last equality are omitted. Thus, $\mathcal{T}_p$ has eigenvalues $1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}$.

It is not difficult to verify that

$$\mathbf{v} = [0\ 1\ 4\ 1\ 0]^T$$

is a (right) 1-eigenvector of $\mathcal{T}_p$. Therefore,

$$v(\omega) = E(\omega)\mathbf{v} = [e^{i2\omega},\ e^{i\omega},\ 1,\ e^{-i\omega},\ e^{-i2\omega}]\mathbf{v} = 4 + 2\cos\omega$$

is a 1-eigenfunction of T_p. ∎

We end this section by introducing the following fundamental lemma for transition operators. This result will be used in the next chapter.

Theorem 3 | **Fundamental lemma of transition operators** | *Let T_p be the transition operator associated with $p = \{p_k\}_{k=0}^N$. Then, for all $f, g \in \mathbb{V}_{2N+1}$,*

$$\int_{-\pi}^{\pi} g(\omega) \left(T_p^n f \right)(\omega) d\omega = \int_{-2^n \pi}^{2^n \pi} g(\omega) \prod_{j=1}^{n} |p\left(\frac{\omega}{2^j}\right)|^2 f\left(\frac{\omega}{2^n}\right) d\omega. \qquad (9.1.12)$$

Proof Since the formula (9.1.12) can be derived by applying mathematical induction, we only give the proof for $n = 1$ in the following. The proof for $n = k + 1$ (using the induction hypothesis; that is, (9.1.12) holds for $n = k$) is similar, and is left as an exercise (see Exercise 9). For $n = 1$, we have

$$\int_{-\pi}^{\pi} g(\omega) \, T_p f(\omega) d\omega = 2 \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} T_p f(2\omega) \, g(2\omega) d\omega$$

$$= 2 \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \left(|p(\omega)|^2 f(\omega) + |p(\omega + \pi)|^2 f(\omega + \pi) \right) g(2\omega) d\omega$$

$$= 2 \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} |p(\omega)|^2 f(\omega) g(2\omega) d\omega + 2 \int_{-\frac{\pi}{2}+\pi}^{\frac{\pi}{2}+\pi} |p(u)|^2 f(u) g(2u - 2\pi) du$$

(substitution $u = \omega + \pi$ was used for the 2nd integral)

$$= 2 \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}+\pi} |p(\omega)|^2 f(\omega) g(2\omega) d\omega \quad \text{(by } g(\omega - 2\pi) = g(\omega))$$

$$= 2 \int_{-\pi}^{\pi} |p(\omega)|^2 f(\omega) g(2\omega) d\omega \quad (|p(\omega)|^2 f(\omega) g(2\omega) \text{ is } 2\pi\text{-periodic)}$$

$$= \int_{-2\pi}^{2\pi} |p\left(\frac{\omega}{2}\right)|^2 f\left(\frac{\omega}{2}\right) g(\omega) d\omega.$$

∎

Exercises

Exercise 1 Let $p = \{p_0, p_1\}$ with $p_0 = p_1 = 1$ be the Haar filter. Compute the eigenvalues of the representation matrix $\mathcal{T}_p$ of the transition operator T_p and the 1-eigenfunctions of T_p, if there are any.

Exercise 2 Repeat Exercise 1 for $p = \{p_k\}_{k=-1}^{1}$ with

$$p_{-1} = p_1 = \frac{1}{2}; \quad p_0 = 1.$$

Exercise 3 Repeat Exercise 1 for $p = \{p_k\}_{k=-2}^{0}$ with

$$p_{-2} = p_0 = \frac{1}{2}; \quad p_{-1} = 1.$$

Exercise 4 Repeat Exercise 1 for $p = \{p_k\}_{k=0}^{3}$ with

$$p_0 = p_3 = \frac{1}{4}; \quad p_1 = p_2 = \frac{3}{4}.$$

Exercise 5 Suppose $p = \{p_k\}_{k=0}^2$ with $p_0 = p_2 = \frac{1}{2}$; $p_1 = 1$. Let $\mathcal{Q}_p$ be the representation matrix of T_p restricted to $\mathbb{V}_3$. Compute the eigenvalues of $\mathcal{Q}_p$. Compare them with the eigenvalues of $\mathcal{T}_p$ obtained in Example 2.

Exercise 6 Repeat Exercise 5 for $p = \{p_k\}_{k=-1}^1$ in Exercise 6.

Exercise 7 Repeat Exercise 5 for $p = \{p_k\}_{k=0}^3$ in Exercise 4.

Exercise 8 Suppose $p = \{p_k\}_{k=0}^N$. Let $\mathcal{T}_p$ and $\mathcal{Q}_p$ be the representation matrices of T_p restricted to $\mathbb{V}_{2N+1}$ and $\mathbb{V}_{2N-1}$, respectively. Find the relation among the eigenvalues of $\mathcal{T}_p$ and $\mathcal{Q}_p$.

Exercise 9 Prove (9.1.12) for $n = k + 1$ under the assumption that (9.1.12) holds for $n = k$.

9.2 Gramian Function $G_\phi(\omega)$

In this section we introduce the Gramian functions $G_\phi(\omega)$ and $G_{\phi,\widetilde{\phi}}(\omega)$. These functions are important for the study of orthogonality, biorthogonality and stability of scaling functions. Some properties of these functions are presented in this section. We also provide the characterizations of orthogonality and stability of scaling functions ϕ in terms of $G_\phi(\omega)$.

For $\phi \in L_2(\mathbb{R})$, define

$$G_\phi(\omega) = \sum_{\ell \in \mathbb{Z}} |\widehat{\phi}(\omega + 2\pi\ell)|^2, \tag{9.2.1}$$

and for $\phi, \widetilde{\phi} \in L_2(\mathbb{R})$, define

$$G_{\widetilde{\phi},\phi}(\omega) = \sum_{\ell \in \mathbb{Z}} \widehat{\widetilde{\phi}}(\omega + 2\pi\ell)\overline{\widehat{\phi}(\omega + 2\pi\ell)}. \tag{9.2.2}$$

Theorem 1 $\boxed{\textbf{Properties of } G_\phi(\omega) \textbf{ and } G_{\widetilde{\phi},\phi}(\omega)}$ *Let ϕ and $\widetilde{\phi}$ be functions in* $L_2(\mathbb{R})$. *Then*:

(i) G_ϕ *and* $G_{\widetilde{\phi},\phi}$ *are 2π-periodic functions in* $L_1[-\pi, \pi]$;
(ii) G_ϕ *and* $G_{\widetilde{\phi},\phi}$ *can be expanded as*

$$G_\phi(\omega) = \sum_k \int_{-\infty}^{\infty} \phi(x)\overline{\phi(x-k)}dx \ e^{-ik\omega}, \tag{9.2.3}$$

$$G_{\widetilde{\phi},\phi}(\omega) = \sum_k \int_{-\infty}^{\infty} \widetilde{\phi}(x)\overline{\phi(x-k)}dx \ e^{-ik\omega}. \tag{9.2.4}$$

Proof To prove (i), we first show $G_\phi \in L_1[-\pi, \pi]$. From Parseval's formula (7.2.7) of Theorem 2 on p.331, we have

$$
\begin{aligned}
2\pi\|\phi\|^2 \ = \ \|\widehat{\phi}\|^2 &= \sum_{\ell=-\infty}^{\infty} \int_{2\pi\ell-\pi}^{2\pi\ell+\pi} \left|\widehat{\phi}(\omega)\right|^2 d\omega \\
&= \sum_{\ell=-\infty}^{\infty} \int_{-\pi}^{\pi} \left|\widehat{\phi}(\omega + 2\pi\ell)\right|^2 d\omega = \int_{-\pi}^{\pi} \sum_{\ell=-\infty}^{\infty} \left|\widehat{\phi}(\omega + 2\pi\ell)\right|^2 d\omega \\
&= \int_{-\pi}^{\pi} G_\phi(\omega) d\omega,
\end{aligned}
$$

where the interchange of the integral and summation signs follows from the fact that $|\widehat{\phi}(\omega + 2\pi\ell)|^2 \geq 0$. Thus $G_\phi(\omega) \in L_1[-\pi, \pi]$.

Using the fact (see Exercise 1) that

$$
|G_{\widetilde{\phi},\phi}(\omega)| \leq \sqrt{G_{\widetilde{\phi}}(\omega)}\sqrt{G_\phi(\omega)}, \tag{9.2.5}
$$

we see $G_{\widetilde{\phi},\phi}(\omega)$ is in $L_1[-\pi, \pi]$. Clearly, $G_\phi(\omega)$ and $G_{\widetilde{\phi},\phi}(\omega)$ are 2π-periodic.

Next, we prove (ii). Since (9.2.3) is a special case of (9.2.4), we only need to show (9.2.4). As a 2π-periodic function in $L_1[-\pi, \pi]$, $G_{\widetilde{\phi},\phi}(\omega)$ can be expanded as its Fourier series which converges in $L_2[-\pi, \pi]$, that is,

$$
G_{\widetilde{\phi},\phi}(\omega) = \sum_k c_k e^{ik\omega} = \sum_k c_{-k} e^{-ik\omega},
$$

where

$$
\begin{aligned}
c_{-k} \ &= \ \frac{1}{2\pi} \int_{-\pi}^{\pi} G_{\widetilde{\phi},\phi}(\omega) e^{ik\omega} d\omega \\
&= \ \frac{1}{2\pi} \sum_{\ell=-\infty}^{\infty} \int_{-\pi}^{\pi} \widehat{\widetilde{\phi}}(\omega + 2\pi\ell)\overline{\widehat{\phi}(\omega + 2\pi\ell)} e^{ik\omega} d\omega \\
&= \ \frac{1}{2\pi} \sum_{\ell=-\infty}^{\infty} \int_{2\pi\ell-\pi}^{2\pi\ell+\pi} \widehat{\widetilde{\phi}}(\omega)\overline{\widehat{\phi}(\omega)} e^{ik\omega} d\omega \\
&= \ \frac{1}{2\pi} \int_{-\infty}^{\infty} \widehat{\widetilde{\phi}}(\omega)\overline{\widehat{\phi}(\omega)} e^{-ik\omega} d\omega \\
&= \ \int_{-\infty}^{\infty} \widetilde{\phi}(x)\overline{\phi(x - k)} dx,
\end{aligned}
$$

where the last equality follows from Parseval's formula, and where we have interchanged the integral and summation signs in the second equality. The justification for the interchanging can be obtained by applying Lebesgue's dominated convergence

theorem to the series of $G_{\widetilde{\phi},\phi}(\omega)$, since the partial sums of this series are bounded by $\sqrt{G_{\widetilde{\phi}}(\omega)}\sqrt{G_\phi(\omega)}$, which is in $L_1[-\pi, \pi]$ (see Exercise 2). ∎

Remark 1 | **Trigonometric polynomials $G_\phi(\omega)$ and $G_{\widetilde{\phi},\phi}(\omega)$** | If $\phi \in L_2(\mathbb{R})$ is compactly supported, then the series in (9.2.3) has finitely many terms, and hence $G_\phi(\omega)$ is a trigonometric polynomial. In particular, if $\mathrm{supp}(\phi) \subseteq [N_1, N_2]$, then

$$G_\phi(\omega) = \sum_{k=-(N_2-N_1-1)}^{N_2-N_1-1} \int_{-\infty}^{\infty} \phi(x)\overline{\phi(x-k)}dx \ e^{-ik\omega}, \qquad (9.2.6)$$

since $\int_{-\infty}^{\infty} \phi(x)\overline{\phi(x-k)}dx = 0$ for $k \geq N_2 - N_1$ or $k \leq -(N_2 - N_1)$.

In the following, for a compactly supported function $\phi \in L_2(\mathbb{R})$, $G_\phi(\omega)$ will denote the trigonometric polynomial given on the right-hand side of (9.2.3). Hence $G_\phi(\omega)$ is continuous and bounded on $\mathbb{R}$.

Similarly, if $\widetilde{\phi}$ and ϕ are compactly supported functions in $L_2(\mathbb{R})$, then $G_{\widetilde{\phi},\phi}(\omega)$ is a trigonometric polynomial. More precisely, if $\mathrm{supp}(\widetilde{\phi}) \subseteq [\widetilde{N}_1, \widetilde{N}_2]$, $\mathrm{supp}(\phi) \subseteq [N_1, N_2]$, then

$$G_{\widetilde{\phi},\phi}(\omega) = \sum_{k=-(N_2-\widetilde{N}_1-1)}^{\widetilde{N}_2-N_1-1} \int_{-\infty}^{\infty} \widetilde{\phi}(x)\overline{\phi(x-k)}dx \ e^{-ik\omega}.$$

∎

Example 1 Let $\phi(x) = \chi_{[0,1)}(x)$. Let us calculate $G_\phi(\omega)$. Since $\mathrm{supp}(\phi) \subseteq [0, 1]$, to find $G_\phi(\omega)$, by (9.2.6), we only need to calculate the integral

$$\int_{-\infty}^{\infty} \phi(x)\overline{\phi(x)}dx = \int_0^1 1^2 dx = 1.$$

Thus $G_\phi(\omega) = 1$ for any $\omega \in \mathbb{R}$. ∎

Example 2 Let $\phi = \min\{x, 2-x\}\chi_{[0,2)}(x)$ be the hat function considered in Examples 1 and 2 in the previous section, namely,

$$\phi(x) = \begin{cases} x, & \text{for } 0 \leq x < 1, \\ 2 - x, & \text{for } 1 \leq x < 2, \\ 0, & \text{elsewhere.} \end{cases}$$

Calculate $G_\phi(\omega)$.

Solution Since $\mathrm{supp}(\phi) \subseteq [0, 2]$, by (9.2.6), to find $G_\phi(\omega)$, we only need to calculate the integrals

$$\int_{-\infty}^{\infty} \phi(x)\phi(x-1)dx = \int_{1}^{2} (2-x)(x-1)dx = \frac{1}{6},$$

$$\int_{-\infty}^{\infty} \phi(x)\phi(x)dx = 2\int_{0}^{1} x^2 dx = \frac{2}{3},$$

$$\int_{-\infty}^{\infty} \phi(x)\phi(x+1)dx = \int_{1}^{2} (2-x)(x-1)dx = \frac{1}{6}.$$

Thus

$$G_{\phi}(\omega) = \sum_{k=-1}^{1} \int_{-\infty}^{\infty} \phi(x)\overline{\phi(x-k)}dx \ e^{-ik\omega}$$

$$= \frac{1}{6}e^{i\omega} + \frac{2}{3} + \frac{1}{6}e^{-i\omega} = \frac{1}{3}(2 + \cos\omega).$$

Recall from Example 2 on p.443 that a 1-eigenfunction $v(\omega)$ of T_p is $v(\omega) = 2 + \cos\omega$. Thus, up to a constant, $G_{\phi}(\omega)$ is equal to $v(\omega)$. The reason behind this is that $G_{\phi}(\omega)$ is a 1-eigenfunction of T_p, as shown in the next theorem. ∎

Theorem 2 $\boxed{G_{\phi}(\omega) \text{ is a 1-eigenfunction of } T_p}$ *Let ϕ be a compactly supported refinable function with the refinement mask $p = \{p_k\}_{k=0}^{N}$. If $\phi \in L^2(\mathbb{R})$, then $G_{\phi}(\omega) \in \mathbb{V}_{2N+1}$ and $G_{\phi}(\omega)$ is a 1-eigenfunction of T_p, that is,*

$$T_p G_{\phi} = G_{\phi}.$$

Proof Since supp$\phi \subseteq [0, N]$ and $\phi \in L_2(\mathbb{R})$, we know, by (9.2.6), $G_{\phi}(\omega) \in \mathbb{V}_{2N-1} \subseteq \mathbb{V}_{2N+1}$. In addition,

$$T_p G_{\phi}(\omega) = |p\left(\frac{\omega}{2}\right)|^2 G_{\phi}\left(\frac{\omega}{2}\right) + |p\left(\frac{\omega}{2}+\pi\right)|^2 G_{\phi}\left(\frac{\omega}{2}+\pi\right)$$

$$= |p\left(\frac{\omega}{2}\right)|^2 \sum_{k} |\widehat{\phi}\left(\frac{\omega}{2}+2k\pi\right)|^2 + |p\left(\frac{\omega}{2}+\pi\right)|^2 \sum_{k} \left|\widehat{\phi}\left(\frac{\omega}{2}+\pi+2k\pi\right)\right|^2$$

$$= \sum_{k} |p\left(\frac{\omega}{2}+2k\pi\right)|^2 |\widehat{\phi}\left(\frac{\omega}{2}+2k\pi\right)|^2$$

$$\quad + \sum_{k} |p\left(\frac{\omega}{2}+\pi+2k\pi\right)|^2 |\widehat{\phi}\left(\frac{\omega}{2}+\pi+2k\pi\right)|^2$$

$$= \sum_{k} |\widehat{\phi}(\omega+4k\pi)|^2 + \sum_{k} |\widehat{\phi}(\omega+2\pi+4k\pi)|^2$$

$$= G_{\phi}(\omega),$$

as desired. ∎

We end this section by showing that the orthogonality and stability of ϕ (where ϕ is not necessarily compactly supported) can be characterized in terms of $G_{\phi}(\omega)$.

Theorem 3 $\boxed{\textbf{Orthogonality of } \phi \textbf{ in terms of } G_\phi(\omega)}$ *Let ϕ be a function in $L_2(\mathbb{R})$. Then ϕ is orthogonal if and only if $G_\phi(\omega) = 1$ a.e. $\omega \in \mathbb{R}$.*

This theorem follows immediately from (9.2.3). Similarly, from (9.2.4), we know two functions $\widetilde{\phi}$ and ϕ in $L_2(\mathbb{R})$ are biorthogonal to each other if and only if $G_{\widetilde{\phi},\phi}(\omega) = 1$ a.e. $\omega \in \mathbb{R}$.

Next we give a characterization on the stability of a function $\phi \in L_2(\mathbb{R})$. Recall that $\phi \in L_2(\mathbb{R})$ is said to be stable if it satisfies (8.2.17) of Sect. 8.2, that is,

$$c \sum_{k \in \mathbb{Z}} |c_k|^2 \le \left\| \sum_{k \in \mathbb{Z}} c_k \phi(x-k) \right\|^2 \le C \sum_{k \in \mathbb{Z}} |c_k|^2, \quad \forall \{c_k\} \in \ell^2(\mathbb{Z}), \qquad (9.2.7)$$

for some constants $c, C > 0$. For a sequence $\{c_k\} \in \ell^2(\mathbb{Z})$, denote $c(\omega) = \sum_k c_k e^{-ik\omega}$. Then $c(\omega) \in L_2[0, 2\pi]$, and

$$\sum_{k \in \mathbb{Z}} |c_k|^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} |c(\omega)|^2 d\omega.$$

By Parseval's formula (7.2.7) of Theorem 2 on p.331, we have

$$2\pi \int_{\mathbb{R}} |\sum_k c_k \phi(x-k)|^2 dx = \int_{\mathbb{R}} |\sum_k c_k e^{-ik\omega} \widehat{\phi}(\omega)|^2 d\omega$$

$$= \int_{\mathbb{R}} |c(\omega) \widehat{\phi}(\omega)|^2 d\omega = \sum_{j \in \mathbb{Z}} \int_{2\pi j - \pi}^{2\pi j + \pi} |c(\omega)|^2 |\widehat{\phi}(\omega)|^2 d\omega$$

$$= \sum_{j \in \mathbb{Z}} \int_{-\pi}^{\pi} |c(\omega)|^2 |\widehat{\phi}(\omega + 2\pi j)|^2 d\omega = \int_{-\pi}^{\pi} |c(\omega)|^2 \sum_{j \in \mathbb{Z}} |\widehat{\phi}(\omega + 2\pi j)|^2 d\omega$$

$$= \int_{-\pi}^{\pi} |c(\omega)|^2 G_\phi(\omega) d\omega.$$

Thus ϕ satisfies (9.2.7) if and only if $0 < c \le G_\phi(\omega) \le C < \infty$, a.e. $\omega \in [-\pi, \pi]$. This conclusion is listed as statement (i) in the following theorem.

Theorem 4 $\boxed{\textbf{Stability of } \phi \textbf{ in terms of } G_\phi(\omega)}$ *Let ϕ be a function in $L_2(\mathbb{R})$. Then*:

(i) *ϕ is stable if and only if $0 < c \le G_\phi(\omega) \le C < \infty$, a.e. $\omega \in [-\pi, \pi]$ for some constants $c, C > 0$.*

(ii) *If, in addition, ϕ is compactly supported, then ϕ is stable if and only if $G_\phi(\omega) > 0$ for $\omega \in [-\pi, \pi]$.*

The statement (ii) in the above theorem follows from (i) and the fact that $G_\phi(\omega)$ is a trigonometric polynomial (see Remark 1). Hence it is bounded and attains its minimum on $[-\pi, \pi]$, which is positive.

Example 3 Let $\phi(x) = \min\{x, 2 - x\}\chi_{[0,2)}(x)$ be the hat function considered in Example 2. Discuss the stability of ϕ by applying Theorem 4.

Solution From Example 2, we have

$$G_\phi(\omega) = \frac{1}{3}(2 + \cos \omega).$$

Since $G_\phi(\omega) > 0$, we conclude, by Theorem 4, that ϕ is stable.   ∎

Example 4 Let $\phi(x) = \frac{1}{3}\chi_{[0,3)}(x)$. Calculate $G_\phi(\omega)$, and then discuss the stability of ϕ.

Solution Since $\text{supp}(\phi) \subseteq [0, 3]$, by (9.2.6), to find $G_\phi(\omega)$, we need to calculate integrals

$$\int_{-\infty}^{\infty} \phi(x)\phi(x - 2)dx = \int_{-\infty}^{\infty} \phi(x)\phi(x + 2)dx = \frac{1}{9},$$

$$\int_{-\infty}^{\infty} \phi(x)\phi(x - 1)dx = \int_{-\infty}^{\infty} \phi(x)\phi(x + 1)dx = \frac{2}{9},$$

$$\int_{-\infty}^{\infty} \phi(x)^2 dx = \frac{1}{3}.$$

Thus

$$\begin{aligned} G_\phi(\omega) &= \sum_{k=-2}^{2} \int_{-\infty}^{\infty} \phi(x)\overline{\phi(x - k)}dx \, e^{-ik\omega} \\ &= \frac{1}{9e^{i2\omega}} + \frac{2}{9e^{i\omega}} + \frac{1}{3} + \frac{2}{9}e^{-i\omega} + \frac{1}{9}e^{-i2\omega} \\ &= \frac{1}{3} + \frac{4}{9}\cos\omega + \frac{2}{9}\cos 2\omega. \end{aligned}$$

Notice that

$$G_\phi\left(\frac{2\pi}{3}\right) = \frac{1}{3} + \frac{4}{9}\cos\frac{2\pi}{3} + \frac{2}{9}\cos\frac{4\pi}{3} = 0.$$

Thus, by Theorem 4, ϕ is not stable.   ∎

Exercises

Exercise 1 Let $\phi, \widetilde{\phi}$ be functions in $L_2(\mathbb{R})$. Show that (9.2.5) holds.

Exercise 2 Suppose $\phi, \widetilde{\phi}$ are functions in $L_2(\mathbb{R})$. For $N = 1, 2, \ldots$, define

$$S_N(\omega) = \sum_{\ell=-N}^{N} \widehat{\widetilde{\phi}}(\omega + 2\pi\ell)\overline{\widehat{\phi}(\omega + 2\pi\ell)}e^{ik\omega},$$

the partial sums of the series $G_{\widetilde{\phi},\phi}(\omega)e^{ik\omega}$, where k is an integer.

(a) Show that $|S_N(\omega)| \leq \sqrt{G_{\widetilde{\phi}}(\omega)}\sqrt{G_\phi(\omega)}$ for $N = 1, 2, \ldots$ and $\omega \in \mathbb{R}$.

(b) Use (a) and Lebesgue's dominated convergence theorem to show that

$$\int_{-\pi}^{\pi} G_{\widetilde{\phi},\phi}(\omega)e^{ik\omega}d\omega = \sum_{\ell=-\infty}^{\infty} \int_{-\pi}^{\pi} \widehat{\widetilde{\phi}}(\omega + 2\pi\ell)\overline{\widehat{\phi}(\omega + 2\pi\ell)}e^{ik\omega}d\omega.$$

Exercise 3 Let $\phi(x) = \chi_{[0,1.5)}(x)$. Calculate $G_\phi(\omega)$, and then discuss the stability of ϕ.

Exercise 4 Repeat Exercise 3 with $\phi(x) = \chi_{[0,1)}(x) + \frac{1}{2}\chi_{[1,2)}(x)$.

Exercise 5 Let $\phi(x)$ be the hat function considered in Example 2. Define $\varphi(x) = \phi(x) + \phi(x-1)$. Calculate $G_\varphi(\omega)$, and then discuss the stability of φ.

Exercise 6 Repeat Exercise 5 with $\varphi(x) = \phi(x) + \frac{1}{2}\phi(x-1)$, where $\phi(x)$ is the hat function.

Exercise 7 Let $\phi = \varphi_0 * \varphi_0 * \varphi_0$ be the quadratic B-spline given by

$$\phi(x) = \begin{cases} \frac{1}{2}x^2, & \text{for } 0 \leq x < 1, \\ -(x-1)^2 + (x-1) + \frac{1}{2}, & \text{for } 1 \leq x < 2, \\ \frac{1}{2}\left(1 - (x-2)\right)^2, & \text{for } 2 \leq x < 3, \\ 0, & \text{elsewhere.} \end{cases}$$

Calculate $G_\phi(\omega)$, and then discuss the stability of ϕ.

9.3 Compactly Supported Orthogonal Wavelets

In this section we consider the construction of compactly supported orthogonal wavelets. The method of construction is based on the architecture of the orthogonal multiresolution approximation (MRA), introduced in Sect. 8.2 of Chap. 8. We will first show that for this MRA approach, the construction of orthogonal wavelets is reduced to constructing orthogonal scaling functions, which is further reduced to the construction of quadrature mirror filter banks (QMF). Next, we discuss the sum-rule properties of lowpass filters and vanishing moments of wavelets. After that we provide the procedures for the construction of orthogonal wavelets and show how these procedures lead to Daubechies orthogonal wavelets.

To construct a (real-valued) wavelet ψ, that is, a function ψ such that $\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$, $j, k \in \mathbb{Z}$ form an orthogonal basis of $L_2(\mathbb{R})$, we start with a real-valued function ϕ which generates an orthogonal MRA $\{\mathbb{V}_j\}$. Let $\mathbb{W}_j = \mathbb{V}_{j+1} \ominus^\perp \mathbb{V}_j$, the orthogonal complement of $\mathbb{V}_j$ in $\mathbb{V}_{j+1}$, namely

$$\mathbb{V}_{j+1} = \mathbb{W}_j \oplus^\perp \mathbb{V}_j, \tag{9.3.1}$$

with $\mathbb{W}_j \perp \mathbb{V}_j$. Then conditions (1°), (2°) and (3°) in the definition of MRA on p.394 imply that $\mathbb{W}_j \perp \mathbb{W}_k$ if $j \neq k$ and $\sum_{j \in \mathbb{Z}} \oplus^\perp \mathbb{W}_j = L^2(\mathbb{R})$ (see Exercise 1).

Observe that, since $\mathbb{V}_j = \{f(x) : f(2^{-j}x) \in \mathbb{V}_0\}$, we have $\mathbb{W}_j = \{g(x) : g(2^{-j}x) \in \mathbb{W}_0\}$ (see Exercise 2). Thus, if the integer translations of a real-valued function ψ form an orthonormal basis of $\mathbb{W}_0$, i.e. $\mathbb{W}_0 = \overline{\text{span}}\{\psi(\cdot - k) : k \in \mathbb{Z}\}$ and

$$\langle \psi, \psi(\cdot - k) \rangle = \delta_k, \;\; k \in \mathbb{Z}, \tag{9.3.2}$$

then, for each j, $\{\psi_{j,k} : k \in \mathbb{Z}\}$ forms an orthonormal basis of $\mathbb{W}_j$, and furthermore, $\{\psi_{j,k} : j, k \in \mathbb{Z}\}$ is an orthonormal basis of $L_2(\mathbb{R})$. Therefore, ψ is an orthogonal wavelet. To construct such a ψ, we notice that, since $\psi \in \mathbb{W}_0 \subseteq \mathbb{V}_1$, we know that ψ is given in terms of $\phi(2x - k)$, $k \in \mathbb{Z}$, that is,

$$\psi(x) = \sum_k q_k \phi(2x - k), \tag{9.3.3}$$

where $\{q_k\}$ is a sequence of real numbers with finitely many q_k nonzero. In addition, we observe that the condition (9.3.1), which can be reduced to the case $j = 0$ only, is equivalent to

$$\langle \phi, \psi(\cdot - k) \rangle = 0, \;\; k \in \mathbb{Z}, \tag{9.3.4}$$

and

$$\phi(2x) = \sum_k c_k \phi(x - k) + \sum_k d_k \psi(x - k), \tag{9.3.5}$$

where c_k, d_k are some real numbers such that the series in (9.3.5) converges in $L_2(\mathbb{R})$. Thus, if a function ψ satisfies the above conditions, then it is an orthogonal wavelet, as stated in the following theorem.

Theorem 1 | **Orth. wavelets derived from orth. scaling functions** | *Let $\phi \in L_2$ ($\mathbb{R}$) be a compactly supported orthogonal refinable function. Let $\{q_k\}$ be a finite sequence such that ψ, given by (9.3.3), satisfies (9.3.2), (9.3.4) and (9.3.5). Then ψ is a compactly supported orthogonal wavelet.*

Example 1 | **Haar wavelet** | Let $\phi(x) = \chi_{[0,1)}(x)$ be the Haar scaling function. Define

$$\psi(x) = \phi(2x) - \phi(2x-1) = \begin{cases} 1, & \text{for } 0 \le x < \frac{1}{2}, \\ -1, & \text{for } \frac{1}{2} \le x < 1, \\ 0, & \text{elsewhere.} \end{cases} \tag{9.3.6}$$

Then we have, for $k \in \mathbb{Z}$,

$$
\begin{aligned}
\langle \psi, \psi(\cdot - k) \rangle &= \langle \phi(2\cdot), \phi(2\cdot -2k) \rangle - \langle \phi(2\cdot), \phi(2\cdot -2k-1) \rangle \\
&\quad - \langle \phi(2\cdot -1), \phi(2\cdot -2k) \rangle + \langle \phi(2\cdot -1), \phi(2\cdot -2k-1) \rangle \\
&= \frac{1}{2}\delta_k - 0 - 0 + \frac{1}{2}\delta_k = \delta_k;
\end{aligned}
$$

and

$$
\begin{aligned}
\langle \phi, \psi(\cdot - k) \rangle &= \langle \phi(2\cdot) + \phi(2\cdot -1), \phi(2\cdot -2k) - \phi(2\cdot -2k-1) \rangle \\
&= \langle \phi(2\cdot), \phi(2\cdot -2k) \rangle - \langle \phi(2\cdot), \phi(2\cdot -2k-1) \rangle \\
&\quad + \langle \phi(2\cdot -1), \phi(2\cdot -2k) \rangle - \langle \phi(2\cdot -1), \phi(2\cdot -2k-1) \rangle \\
&= \frac{1}{2}\delta_k - 0 + 0 - \frac{1}{2}\delta_k = 0.
\end{aligned}
$$

In addition, we have

$$\phi(2x) = \frac{1}{2}\phi(x) + \frac{1}{2}\psi(x),$$

that is, (9.3.5) is satisfied. Thus, by Theorem 1, ψ is an orthogonal wavelet, which is called the **Haar wavelet**.

On the other hand, one can show directly that $\{\psi_{j,k}(x)\}$, $j, k \in \mathbb{Z}$, is complete in $L_2(\mathbb{R})$ and orthogonal (see Exercise 3). ∎

Next, we discuss how to apply Theorem 1 to construct orthogonal wavelets. The idea is to start with the refinement mask $\{p_k\}$ for an orthogonal scaling function ϕ, and then we derive the sequence $\{q_k\}$ for the orthogonal wavelet ψ directly from $\{p_k\}$. Here, we focus on compactly supported wavelets, so that $\{p_k\}$ and $\{q_k\}$ are finite sequences. Let $p(\omega)$ and $q(\omega)$ be the associated two-scale symbols

$$p(\omega) = \frac{1}{2}\sum_k p_k e^{-ik\omega}, \quad q(\omega) = \frac{1}{2}\sum_k q_k e^{-ik\omega},$$

which are trigonometric polynomials.

Theorem 2 $\boxed{\text{Orthogonality implies QMF}}$ *Let $\phi \in L_2(\mathbb{R})$ be a refinable function with refinement mask $\{p_k\}$, and suppose ϕ is orthogonal. Then $p(\omega)$ is a QMF, that is, $p(\omega)$ satisfies*

$$|p(\omega)|^2 + |p(\omega + \pi)|^2 = 1, \quad \omega \in \mathbb{R}. \tag{9.3.7}$$

Furthermore, ψ, defined by (9.3.3), satisfies (9.3.2) and (9.3.4) if and only if

$$|q(\omega)|^2 + |q(\omega + \pi)|^2 = 1, \ \omega \in \mathbb{R}, \tag{9.3.8}$$

$$p(\omega)\overline{q(\omega)} + p(\omega + \pi)\overline{q(\omega + \pi)} = 0, \ \omega \in \mathbb{R}. \tag{9.3.9}$$

Proof By the refinement of ϕ, we have

$$
\begin{aligned}
G_\phi(2\omega) &= \sum_{\ell \in \mathbb{Z}} \left| \widehat{\phi}(2\omega + 2\pi\ell) \right|^2 = \sum_{\ell \in \mathbb{Z}} \left| p(\omega + \pi\ell)\widehat{\phi}(\omega + \pi\ell) \right|^2 \\
&= \sum_{k \in \mathbb{Z}} \left| p(\omega + 2k\pi)\widehat{\phi}(\omega + 2k\pi) \right|^2 + \sum_{k \in \mathbb{Z}} \left| p(\omega + (2k+1)\pi)\widehat{\phi}(\omega + (2k+1)\pi) \right|^2 \\
&= |p(\omega)|^2 G_\phi(\omega) + |p(\omega + \pi)|^2 G_\phi(\omega + \pi).
\end{aligned}
\tag{9.3.10}
$$

Thus, if ϕ is orthogonal, or equivalently, $G_\phi(\omega) = 1$, then we can deduce from (9.3.10) that $p(\omega)$ is a QMF.

Similarly, we have

$$
\begin{aligned}
G_\psi(2\omega) &= |q(\omega)|^2 G_\phi(\omega) + |q(\omega + \pi)|^2 G_\phi(\omega + \pi), \\
G_{\phi,\psi}(2\omega) &= p(\omega)\overline{q(\omega)} G_\phi(\omega) + p(\omega + \pi)\overline{q(\omega + \pi)} G_\phi(\omega + \pi).
\end{aligned}
$$

Thus, under the assumption that $G_\phi(\omega) = 1$, ψ satisfies (9.3.2) and (9.3.4), which are equivalent to

$$G_\psi(\omega) = 1 \quad \text{and} \quad G_{\phi,\psi}(\omega) = 0,$$

if and only if p and q satisfy (9.3.8) and (9.3.9). ∎

Following the remark in Sect. 8.2 of Chap. 8, on p.399, for a QMF $p(\omega)$, we may simply choose the highpass filter $q(\omega)$ for the orthogonal wavelet to be

$$q(\omega) = -\overline{p(\omega + \pi)} e^{-i(2L-1)\omega} = -p(\pi - \omega) e^{-i(2L-1)\omega}, \tag{9.3.11}$$

or equivalently, $q_k = (-1)^k p_{2L-1-k}, k \in \mathbb{Z}$, where L is an integer. Then $q(\omega)$ satisfies (9.3.8) and (9.3.9).

Next, we provide a characterization for the orthogonality of ϕ in terms of its refinement mask p. More precisely, the characterization is given by certain conditions on the eigenvalues of the transition operator T_p associated with p.

Recall that, for $p = \{p_k\}$ supported in $[N_1, N_2]$, the transition operator T_p associated with p is the linear operator defined on $\mathbb{V}_{2(N_2-N_1)+1}$ by

$$(T_p)v(\omega) = \left| p\left(\frac{\omega}{2}\right) \right|^2 v\left(\frac{\omega}{2}\right) + \left| p\left(\frac{\omega}{2} + \pi\right) \right|^2 v\left(\frac{\omega}{2} + \pi\right), \ v(\omega) \in \mathbb{V}_{2(N_2-N_1)+1},$$

where $\mathbb{V}_{2(N_2-N_1)+1}$ denotes the linear span of $e^{-ik\omega}$, $|k| \leq N_2 - N_1$, that is,

$$\mathbb{V}_{2(N_2-N_1)+1} = \{ \sum_{k=-(N_2-N_1)}^{N_2-N_1} c_k \, e^{-ik\omega} \; : \; c_k \in \mathbb{C} \}.$$

Then $T_p v(\omega) \in \mathbb{V}_{2(N_2-N_1)+1}$ for any $v(\omega) \in \mathbb{V}_{2(N_2-N_1)+1}$.

Definition 1 $\boxed{\textbf{Condition E}}$ *A matrix or a linear operator T on a finite-dimensional space is said to satisfy* **Condition E** *if 1 is a simple eigenvalue of T, and any other eigenvalue λ of T satisfies $|\lambda| < 1$.*

Theorem 3 $\boxed{\textbf{Characterization of orthogonality of refinable functions}}$ *Suppose a trigonometric polynomial $p(\omega)$ is a QMF with $p(0) = 1$. Then the associated refinable function ϕ is in $L_2(\mathbb{R})$ and orthogonal if and only if $p(\pi) = 0$ and the transition operator T_p satisfies Condition* E.

The proof of Theorem 3 is provided in Sect. 10.2 of Chap. 10, on p. 504.

Theorem 4 $\boxed{\textbf{Orthogonal wavelets derived from QMFs}}$ *Let the trigonometric polynomial $p(\omega)$ be a QMF with $p(0) = 1$, $p(\pi) = 0$, and ϕ be the refinable function associated with p. Let ψ be the function given by (9.3.3), with $\{q_k\}$ defined by (9.3.11). If the transition operator T_p associated with p satisfies Condition* E, *then ψ is a compactly supported orthogonal wavelet.*

Proof By Theorem 3, ϕ is in $L_2(\mathbb{R})$ and orthogonal. Thus, by Theorem 2, ϕ and ψ satisfy (9.3.2) and (9.3.4), since $q(\omega)$ satisfies (9.3.8)–(9.3.9). Therefore, by Theorem 1, to prove Theorem 4, namely to show ψ to be a compactly supported orthogonal wavelet, we only need to verify (9.3.5), which is carried out below.

Let $M_{p,q}(\omega)$ be the modulation matrix of $p(\omega)$, $q(\omega)$ as defined in Sect. 8.3 of Chap. 8, that is,

$$M_{p,q}(\omega) = \begin{bmatrix} p(\omega) & p(\omega + \pi) \\ q(\omega) & q(\omega + \pi) \end{bmatrix}. \tag{9.3.12}$$

Then (9.3.7)–(9.3.9) is equivalent to

$$\overline{M_{p,q}(\omega)} M_{p,q}(\omega)^T = I_2, \; \omega \in \mathbb{R},$$

or

$$M_{p,q}(\omega)^T \overline{M_{p,q}(\omega)} = I_2, \; \omega \in \mathbb{R}.$$

In particular, we have

$$\overline{p(\omega)} p(\omega) + \overline{q(\omega)} q(\omega) = 1, \; \overline{p(\omega)} p(\omega + \pi) + \overline{q(\omega)} q(\omega + \pi) = 0,$$

so that

$$\left(\overline{p(\omega)} + \overline{p(\omega + \pi)}\right)p(\omega) + \left(\overline{q(\omega)} + \overline{q(\omega + \pi)}\right)q(\omega) = 1.$$

Hence

$$\widehat{\phi}\left(\frac{\omega}{2}\right) = \left(\overline{p\left(\frac{\omega}{2}\right)} + \overline{p\left(\frac{\omega}{2} + \pi\right)}\right)p\left(\frac{\omega}{2}\right)\widehat{\phi}\left(\frac{\omega}{2}\right) + \left(\overline{q\left(\frac{\omega}{2}\right)} + \overline{q\left(\frac{\omega}{2} + \pi\right)}\right)q\left(\frac{\omega}{2}\right)\widehat{\phi}\left(\frac{\omega}{2}\right)$$

$$= \frac{1}{2}\sum_k \left(1 + (-1)^k\right)\overline{p_k}e^{ik\frac{\omega}{2}}\,\widehat{\phi}(\omega) + \frac{1}{2}\sum_k \left(1 + (-1)^k\right)\overline{q_k}e^{ik\frac{\omega}{2}}\,\widehat{\psi}(\omega)$$

$$= \sum_n p_{2n}e^{in\omega}\,\widehat{\phi}(\omega) + \sum_n q_{2n}e^{in\omega}\,\widehat{\psi}(\omega),$$

where, in the last equality, we use the assumption that p_k and q_k are real numbers. Thus

$$\phi(2x) = \frac{1}{2}\sum_n p_{-2n}\phi(x - n) + \frac{1}{2}\sum_n q_{-2n}\psi(x - n).$$

Hence we have shown that (9.3.5) holds. ∎

Definition 2 | **Sum-rule order** | *A finite sequence* $p = \{p_k\}$ *is said to have* **sum-rule order** L *if L is the largest integer, for which the two-scale symbol $p(\omega)$ satisfies*

$$p(0) = 1, \quad \frac{d^\ell}{d\omega^\ell}p(\pi) = 0, \text{ for } \ell = 0, 1, \ldots, L - 1, \quad (9.3.13)$$

or equivalently

$$\sum_k p_k = 2, \quad \sum_k (2k)^\ell p_{2k} = \sum_k (2k+1)^\ell p_{2k+1}, \text{ for } \ell = 0, 1, \ldots, L-1. \quad (9.3.14)$$

In this case, we also say that $p(\omega)$ has sum-rule order L.

A (compactly supported) function $\psi \in L_2(\mathbb{R})$ is said to have **vanishing moments of order** L if

$$\int_{-\infty}^{\infty} \psi(x)x^\ell dx = 0, \text{ for } \ell = 0, 1, \ldots, L - 1.$$

Let $q(\omega)$ be the trigonometric polynomial defined by (9.3.11). Then

$$\frac{d^\ell}{d\omega^\ell}q(\omega) = \frac{d^\ell}{d\omega^\ell}\left(-p(\pi - \omega)e^{-i(2L-1)\omega}\right)$$

$$= -\sum_{j=0}^{\ell}\binom{\ell}{j}\frac{d^j}{d\omega^j}p(\pi - \omega)(-1)^j\frac{d^{\ell-j}}{d\omega^{\ell-j}}\left(e^{-i(2L-1)\omega}\right).$$

Thus, if $p(\omega)$ has sum-rule order L, then

$$\frac{d^\ell}{d\omega^\ell}q(0) = 0, \text{ for } 0 \leq \ell \leq L - 1. \tag{9.3.15}$$

Condition (9.3.15) can be written as

$$\sum_k k^\ell q_k = 0, \text{ for } 0 \leq \ell \leq L - 1 \tag{9.3.16}$$

(see Exercise 14). This property of $\{q_k\}$ is called the discrete polynomial annihilation.

Theorem 5 $\boxed{\textbf{Sum-rules of QMF implies vanishing moments}}$ *Let ϕ be an orthogonal refinable function associated with a trigonometric polynomial $p(\omega)$. Let ψ be the function given by (9.3.3), with $\{q_k\}$ defined by (9.3.11). If $p(\omega)$ has sum-rule order L, then ψ has vanishing moments of order L.*

Proof Since ϕ, ψ are compactly supported, we know $\widehat{\phi}, \widehat{\psi} \in C^\infty$, by the property (v) of Fourier transform in Theorem 1 on p.320. Thus, from the fact that

$$\frac{d^\ell}{d\omega^\ell}\widehat{\psi}(\omega) = \frac{d^\ell}{d\omega^\ell}\left(q\left(\frac{\omega}{2}\right)\widehat{\phi}\left(\frac{\omega}{2}\right)\right)$$

$$= \sum_{j=0}^{\ell} \binom{\ell}{j}\frac{d^j}{d\omega^j}q\left(\frac{\omega}{2}\right)\left(\frac{1}{2}\right)^j\frac{d^{\ell-j}}{d\omega^{\ell-j}}\left(\widehat{\phi}\left(\frac{\omega}{2}\right)\right),$$

and using also (9.3.15), we conclude that

$$\frac{d^\ell}{d\omega^\ell}\widehat{\psi}(0) = 0, \text{ for } 0 \leq \ell \leq L - 1.$$

This, together with (7.1.2) on p.320, yields

$$\int_{-\infty}^{\infty} x^\ell \psi(x)dx = i^k\frac{d^\ell}{d\omega^\ell}\widehat{\psi}(0) = 0, \text{ for } 0 \leq \ell \leq L - 1,$$

that is, ψ has vanishing moments of order L. $\blacksquare$

In fact, one can show that if $p(\omega), q(\omega)$ satisfy (9.3.9), and $p(\omega)$ has sum-rule order L, then $q(\omega)$ satisfies (9.3.15) (see Exercise 13). Hence the corresponding orthogonal wavelet ψ has vanishing moments of order L if the scaling function ϕ is in $L_2(\mathbb{R})$.

$\boxed{\textbf{Construction of compactly supported orthogonal wavelets}}$ By Theorem 5, to construct an orthogonal wavelet with vanishing moments of order L, we start with a QMF $p(\omega)$ of the form

$$p(\omega) = \left(\frac{1 + e^{-i\omega}}{2}\right)^L p_0(\omega), \tag{9.3.17}$$

where $p_0(\omega)$ is a trigonometric polynomial. For such a $p(\omega)$, we have

$$|p(\omega)|^2 = \left(\cos\frac{\omega}{2}\right)^{2L} |p_0(\omega)|^2.$$

Since the coefficients p_k of $p(\omega)$ are real numbers, we have

$$|p_0(\omega)|^2 = p_0(\omega)\overline{p_0(\omega)} = \overline{p_0(-\omega)}p_0(-\omega) = |p_0(-\omega)|^2,$$

that is, $|p_0(\omega)|^2$ is an even trigonometric polynomial, and it can be written as a function of $\cos\omega$. Therefore, with $\cos\omega = 1 - 2\sin^2\frac{\omega}{2}$, $|p_0(\omega)|^2$ can be expressed as

$$|p_0(\omega)|^2 = P\left(\sin^2\frac{\omega}{2}\right),$$

where P is a polynomial. Hence, we conclude that $p(\omega)$ in (9.3.17) is a QMF, namely, it satisfies (9.3.7), if and only if P satisfies

$$\cos^{2L}\frac{\omega}{2} P\left(\sin^2\frac{\omega}{2}\right) + \sin^{2L}\frac{\omega}{2}P\left(\cos^2\frac{\omega}{2}\right) = 1, \ \omega \in \mathbb{R},$$

or equivalently, with $y = \sin^2\frac{\omega}{2}$, P satisfies

$$(1 - y)^L P(y) + y^L P(1 - y) = 1, \ y \in [0, 1]. \tag{9.3.18}$$

Let P_{L-1} be the polynomial defined by

$$P_{L-1}(y) = \sum_{k=0}^{L-1} \binom{L-1+k}{k} y^k. \tag{9.3.19}$$

Then $P(y) = P_{L-1}(y) + y^L R\left(\frac{1}{2} - y\right)$ satisfies (9.3.18), where R is an odd polynomial (including the zero polynomial). In the following, we list $P_{L-1}(y)$ for $1 \le L \le 5$:

$$\begin{aligned}
P_0(y) &= 1, \\
P_1(y) &= 1 + 2y, \\
P_2(y) &= 1 + 3y + 6y^2, \\
P_3(y) &= 1 + 4y + 10y^2 + 20y^3, \\
P_4(y) &= 1 + 5y + 15y^2 + 35y^3 + 70y^4.
\end{aligned} \tag{9.3.20}$$

By factorizing $P_{L-1}(\sin^2\frac{\omega}{2})$ into $p_0(\omega)\overline{p_0(\omega)}$, we obtain a QMF $p(\omega)$ with sum-rule order L. This resulting QMF $p(\omega)$ has $2L$ nonzero coefficients p_k, and it is called the $2L$-tap orthogonal filter or $2L$-tap Daubechies orthogonal filter (the D_{2L} filter for short). In the following, we look at the cases with $L \le 4$. ∎

As before, we use the notation

$$z = e^{-i\omega}.$$

With this notation, we have

$$\cos^2\frac{\omega}{2} = \frac{1}{4}(1+z)\left(1+\frac{1}{z}\right), \ \sin^2\frac{\omega}{2} = \frac{1}{4}\left(2 - z - \frac{1}{z}\right).$$

Haar wavelet For $L = 1$, since $P_0(y) = 1$, we have

$$|p(\omega)|^2 = \cos^2\frac{\omega}{2} = \frac{1}{4}(1+z)\left(1+\frac{1}{z}\right).$$

Thus

$$p(\omega) = \frac{1}{2}(1+z) = \frac{1}{2}(1+e^{-i\omega}),$$

and

$$q(\omega) = -\overline{p(\omega+\pi)}e^{-i(2\times 1-1)\omega} = \frac{1}{2}(1 - e^{-i\omega}).$$

In this case, we reach the Haar wavelet. ∎

D_4 filter For $L = 2$, with $y = \sin^2\frac{\omega}{2} = \frac{1}{4}(2 - z - \frac{1}{z})$, we have

$$P_1(y) = 1 + 2y = 2 - \frac{1}{2}z - \frac{1}{2}\frac{1}{z} = -\frac{1}{2z}(z^2 - 4z + 1).$$

Notice that the roots of

$$z^2 - 4z + 1 = 0$$

are $r_1 = 2 - \sqrt{3}, r_2 = 2 + \sqrt{3}$, with $r_2 = \frac{1}{r_1}$. Thus

$$P_1(y) = -\frac{1}{2z}(z - r_1)(z - r_2) = -\frac{1}{2zr_1}(z - r_1)(r_1 z - r_1 r_2)$$

$$= \frac{1}{2r_1}(z - r_1)\left(\frac{1}{z} - r_1\right).$$

Therefore, we may choose

$$p_0(\omega) = \frac{1}{\sqrt{2r_1}}(z - r_1) = \frac{1 + \sqrt{3}}{2}\left(z - (2 - \sqrt{3})\right),$$

and we reach the D_4 orthogonal filter:

$$
\begin{aligned}
p(\omega) &= \left(\frac{1+z}{2}\right)^2 p_0(\omega) = \left(\frac{1+z}{2}\right)^2 \frac{1+\sqrt{3}}{2}\left(z - (2 - \sqrt{3})\right) \\
&= \frac{1 - \sqrt{3}}{8} + \frac{3 - \sqrt{3}}{8}z + \frac{3 + \sqrt{3}}{8}z^2 + \frac{1 + \sqrt{3}}{8}z^3. \tag{9.3.21}
\end{aligned}
$$

The corresponding highpass filter $q(\omega)$ for the orthogonal wavelet is given by

$$
\begin{aligned}
q(\omega) &= -\overline{p(\omega + \pi)}e^{-i(2\times2-1)\omega} \\
&= \frac{1 + \sqrt{3}}{8} - \frac{3 + \sqrt{3}}{8}z + \frac{3 - \sqrt{3}}{8}z^2 - \frac{1 - \sqrt{3}}{8}z^3.
\end{aligned}
$$

The graphs of the corresponding scaling function ϕ and orthogonal wavelet ψ are displayed in Fig. 8.2 in Sect. 8.2 of Chap. 8, on p.400.

Observe that $P_1(y)$ can also be factorized as

$$P_1(y) = \frac{1}{2r_2}(z - r_2)\left(\frac{1}{z} - r_2\right) = \frac{1}{2r_2}(r_2 - z)\left(r_2 - \frac{1}{z}\right).$$

Thus, one may choose

$$p_0(\omega) = \frac{1}{\sqrt{2r_2}}(r_2 - z) = \frac{\sqrt{3} - 1}{2}\left(2 + \sqrt{3} - z\right).$$

In this case, the corresponding 4-tap orthogonal filter, denoted by $r(\omega)$, is

$$
\begin{aligned}
r(\omega) &= \left(\frac{1+z}{2}\right)^2 \frac{\sqrt{3} - 1}{2}(2 + \sqrt{3} - z) \\
&= \frac{1 + \sqrt{3}}{8} + \frac{3 + \sqrt{3}}{8}z + \frac{3 - \sqrt{3}}{8}z^2 + \frac{1 - \sqrt{3}}{8}z^3.
\end{aligned}
$$

This lowpass filter $r(\omega)$ is $p(-\omega)e^{-i4\omega}$, where $p(\omega)$ is the 4-tap orthogonal filter given by (9.3.21). One can easily show that if $p(\omega)$ is a QMF, then $p(-\omega)e^{-in\omega}$ is also a QMF, where $n \in \mathbb{N}$ (see Exercise 4).   ■

$\boxed{D_6 \text{ filter}}$ For $L = 3$, we first factor $P_2(y) = 1 + 3y + 6y^2$ as

$$P_2(y) = 6(y - y_1)(y - y_2),$$

where $y_1, y_2 = -\frac{1}{4} \pm i\frac{\sqrt{15}}{12}$. With $y = \sin^2\frac{\omega}{2} = \frac{1}{4}(2 - z - \frac{1}{z})$, we have

$$P_2(y) = \frac{3}{8}\left(z + \frac{1}{z} - c\right)\left(z + \frac{1}{z} - \bar{c}\right),$$

where $c = 3 - i\frac{\sqrt{15}}{3}$. Let

$$z_0 = \frac{1}{2}\left(c - \sqrt{c^2 - 4}\right) = \frac{3}{2} - \frac{\sqrt{12\sqrt{10} + 15}}{6} - \frac{\sqrt{15} - \sqrt{12\sqrt{10} - 15}}{6}i$$

be a root of $z^2 - cz + 1 = 0$. Then $P_2(y)$ is further factorized as

$$\begin{aligned}
P_2(y) &= \frac{3}{8}\frac{1}{z^2}(z^2 - cz + 1)(z^2 - \bar{c}z + 1) \\
&= \frac{3}{8}\frac{1}{z^2}(z - z_0)\left(z - \frac{1}{z_0}\right)(z - \bar{z}_0)\left(z - \frac{1}{\bar{z}_0}\right) \\
&= \frac{3}{8}(z - z_0)(z - \bar{z}_0)\left(1 - \frac{1}{zz_0}\right)\left(1 - \frac{1}{z\bar{z}_0}\right) \\
&= \frac{3}{8}\frac{1}{|z_0|^2}(z - z_0)(z - \bar{z}_0)\left(\frac{1}{z} - z_0\right)\left(\frac{1}{z} - \bar{z}_0\right).
\end{aligned}$$

Thus, we have the corresponding QMF $p(\omega)$, given by

$$\begin{aligned}
p(\omega) &= \left(\frac{1+z}{2}\right)^3\sqrt{\frac{3}{8}\frac{1}{|z_0|}}(z - z_0)(z - \bar{z}_0) \\
&= \left(\frac{1+z}{2}\right)^3\frac{\sqrt{6}}{4}\frac{1}{|z_0|}\left(z^2 - 2\,Re(z_0)\,z + |z_0|^2\right) \\
&= \left(\frac{1+z}{2}\right)^3\left(\frac{1 + \sqrt{10} - \sqrt{5 + 2\sqrt{10}}}{4} + \frac{1 - \sqrt{10}}{2}z + \frac{1 + \sqrt{10} + \sqrt{5 + 2\sqrt{10}}}{4}z^2\right),
\end{aligned}$$

where the nonzero p_k are given by

$$p_0 = \frac{1}{16}\left(1 + \sqrt{10} - \sqrt{5 + 2\sqrt{10}}\right),$$

$$p_1 = \frac{1}{16}\left(5 + \sqrt{10} - 3\sqrt{5 + 2\sqrt{10}}\right),$$

$$p_2 = \frac{1}{8}\left(5 - \sqrt{10} - \sqrt{5 + 2\sqrt{10}}\right),$$
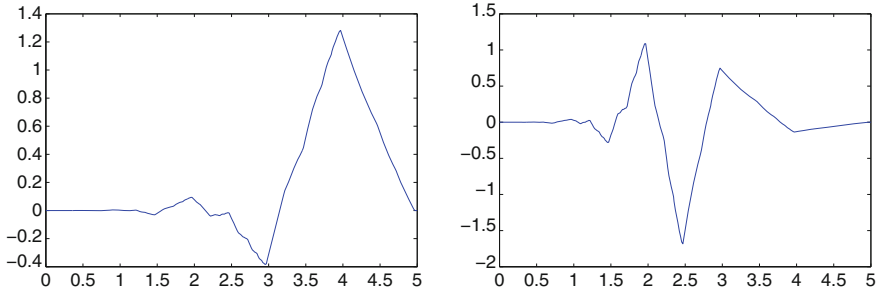
Fig. 9.1. D_6 scaling function ϕ (on *left*) and wavelet ψ (on *right*)

$$p_3 = \frac{1}{8}\left(5 - \sqrt{10} + \sqrt{5 + 2\sqrt{10}}\right),$$

$$p_4 = \frac{1}{16}\left(5 + \sqrt{10} + 3\sqrt{5 + 2\sqrt{10}}\right),$$

$$p_5 = \frac{1}{16}\left(1 + \sqrt{10} + \sqrt{5 + 2\sqrt{10}}\right).$$

The highpass filter $\{q_k\}$ can be chosen as

$$q_k = (-1)^k p_{5-k}, \ k \in \mathbb{Z}.$$

See Fig. 9.1 for the scaling function ϕ and orthogonal wavelet ψ. ∎

$\boxed{D_8 \text{ filter}}$ For $L = 4$, $P_3(y) = 1 + 4y + 10y^2 + 20y^3$ can be factorized as

$$P_3(y) \qquad = 20(y - \theta)\left(y^2 + \left(\theta + \frac{1}{2}\right)y - \frac{1}{20\theta}\right) \qquad (9.3.22)$$
$$= 20(y - \theta)(y - \lambda)(y - \bar{\lambda}),$$

where

$$\theta = -\frac{1}{6} + \frac{1}{30}\left(\sqrt[3]{105\sqrt{15} - 350} - \sqrt[3]{105\sqrt{15} + 350}\right)$$
$$\approx -0.3423840948583691,$$
$$\lambda \approx -0.0788079525708154 + 0.3739306454336101\, i. \qquad (9.3.23)$$

Thus, with $y = \frac{1}{4}(2 - z - \frac{1}{z})$, we can factorize $P_3(y)$ as follows, as we did for the cases $L = 1, 2$:

$$P_3(y) = -\frac{5}{16z^3}\left(z^2 - (2-4\theta)z + 1\right)\left(z^2 - (2-4\lambda)z + 1\right)\left(z^2 - (2-4\bar{\lambda})z + 1\right)$$

$$= -\frac{5}{16z^3}(z - z_0)(z - \frac{1}{z_0})(z - z_1)(z - \frac{1}{z_1})(z - \overline{z_1})(z - \frac{1}{\overline{z_1}})$$

$$= \frac{5}{16z_0|z_1|^2}(z - z_0)\left(\frac{1}{z} - z_0\right)(z - z_1)(z - \overline{z_1})\left(\frac{1}{z} - \overline{z_1}\right)\left(\frac{1}{z} - z_1\right),$$

where

$$z_0 \approx 0.32887591778603087,$$
$$z_1 \approx 0.28409629819182162 + 0.24322822591037987\, i.$$

Therefore, with

$$p_0(\omega) = \frac{\sqrt{5}}{4}\frac{1}{\sqrt{z_0}}\frac{1}{|z_1|}(z - z_0)(z - z_1)(z - \overline{z_1}),$$

we have a QMF with sum-rule order 4, given by

$$p(\omega) = \left(\frac{1+z}{2}\right)^3 p_0(\omega),$$

with 8 nonzero p_k given by

$$p_0 = -0.0149869893303615, \quad p_1 = 0.0465036010709818,$$
$$p_2 = 0.0436163004741773, \quad p_3 = -0.2645071673690398,$$
$$p_4 = -0.0395750262356446, \quad p_5 = 0.8922001382467596,$$
$$p_6 = 1.0109457150918289, \quad p_7 = 0.3258034280512984.$$

The corresponding highpass filter $\{q_k\}$ can be chosen as

$$q_k = (-1)^k p_{7-k}, \ k \in \mathbb{Z}.$$

See Fig. 9.2 for the scaling function ϕ and orthogonal wavelet ψ. ∎

Observe that the orthogonal wavelets of the D_4, D_6, D_8 QMFs have vanishing moments of orders 2, 3, 4, respectively. In this way, one can construct orthogonal wavelets with higher vanishing moments.

Exercises

Exercise 1 Suppose $\{\mathbb{V}_j\}$ is an orthogonal MRA. Let $\mathbb{W}_j$ be the orthogonal complement of $\mathbb{V}_j$ in $\mathbb{V}_{j+1}$. Show that $\mathbb{W}_j \perp \mathbb{W}_k$ if $j \neq k$ and $\sum_{j\in\mathbb{Z}} \oplus \mathbb{W}_j = L_2(\mathbb{R})$.
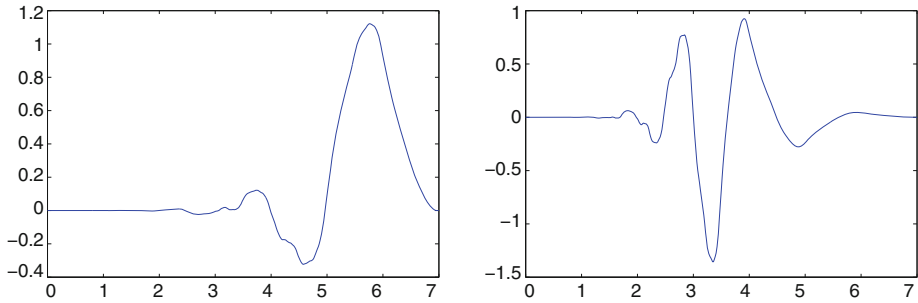
Fig. 9.2. D_8 scaling function ϕ (on *left*) and wavelet ψ (on *right*)

Exercise 2 Suppose $\{\mathbb{V}_j\}$ is an orthogonal MRA. Let $\mathbb{W}_j$ be the orthogonal complement of $\mathbb{V}_j$ in $\mathbb{V}_{j+1}$. Show that $\mathbb{W}_j = \{g(x) : g(2^{-j}x) \in \mathbb{W}_0\}$.

Exercise 3 Let ψ be the Haar wavelet defined by (9.3.6) in Example 1. Show directly that $\{\psi_{j,k}(x)\}$, $j, k \in \mathbb{Z}$ is orthogonal.

Exercise 4 Show that if $p(\omega)$ is a QMF, then $p(-\omega)e^{-in\omega}$ is also a QMF, where $n \in \mathbb{Z}$.

Exercise 5 Suppose FIR filters $p(\omega), q(\omega)$ satisfy (9.3.7)–(9.3.9). Show that, for any $n, m \in \mathbb{Z}$, $p(\omega)e^{-in\omega}, q(\omega)e^{-i(n+2m)\omega}$ also satisfy (9.3.7)–(9.3.9).

Exercise 6 Suppose $\phi(x)$ is the refinable function with two-scale symbol $p(\omega)$. Let $\varphi(x)$ be the refinable function with two-scale symbol $p(\omega)e^{-in\omega}$, where $n \in \mathbb{Z}$. Find the relation between $\phi(x)$ and $\varphi(x)$.

Exercise 7 Let $\psi(x)$ be the function defined by (9.3.3), with a finite sequence $\{q_k\}$. Suppose $f(x)$ is the function defined by (9.3.3), with two-scale symbol $q(\omega)e^{-i2m\omega}$, where $m \in \mathbb{Z}$. Find the relation between $\psi(x)$ and $f(x)$.

Exercise 8 Show that (9.3.13) is equivalent to (9.3.14).

Exercise 9 Let $p = \{p_k\}$ be a refinement mask with nonzero p_k given by

$$p_0 = p_1 = p_2 = p_3 = \frac{1}{2}.$$

Find the sum-rule order of p.

Exercise 10 Repeat Exercise 9 for $p = \{p_k\}$ with nonzero p_k given by

$$p_0 = \frac{1}{4}, \quad p_1 = \frac{3}{4}, \quad p_2 = \frac{3}{4}, \quad p_3 = \frac{1}{4}.$$

Exercise 11 Repeat Exercise 9 for $p = \{p_k\}$ with nonzero p_k given by

$$p_{-1} = \frac{1}{4}, \quad p_0 = \frac{3}{4}, \quad p_1 = \frac{3}{4}, \quad p_2 = \frac{1}{4}.$$

Exercise 12 Suppose the FIR filter $p(\omega)$ has sum-rule order L. Show that, for any $n \in \mathbb{Z}$, $p(\omega)e^{-in\omega}$ also has sum-rule order L.

Exercise 13 Suppose FIR filters $p(\omega), q(\omega)$ satisfy (9.3.9). Show that if $p(\omega)$ has sum-rule order L, then $q(\omega)$ satisfies (9.3.15).

   Hint: Differentiate both sides of (9.3.9) with respect to ω and set $\omega = 0$.

Exercise 14 Show that (9.3.15) is equivalent to (9.3.16).

Exercise 15 Let $P(y) = P_{L-1}(y)$ be the polynomial defined by (9.3.20). For $L = 2, 3, 4$, show that $P(y)$ satisfies (9.3.18).

Exercise 16   Let $p(\omega)$ be the orthogonal filter given by (9.3.21). Show that $p(\omega)$ has sum-rule order 2.

9.4   Compactly Supported Biorthogonal Wavelets

The compactly supported orthogonal wavelets discussed in the previous section lack symmetry. In many applications, symmetry is an important property. For example, in application to image compression, to deal with the boundary extension of the image, symmetric filters are desirable. On the other hand, expanding a signal in a Riesz basis is enough in some applications. Relaxing the orthogonality condition will result in symmetric wavelets. In this section, we discuss the construction of compactly supported biorthogonal wavelets with symmetry.

   In the beginning of this section, we introduce the definitions of Bessel sequences and Riesz bases, after which we show how MRAs lead to biorthogonal wavelets, which is reduced to the construction of biorthogonal filters. Next, we discuss the symmetry of biorthogonal wavelets. Finally, we provide the procedures of the construction of biorthogonal wavelets, and show how these procedures lead to some biorthogonal wavelets, including 5/3-tap and 9/7-tap biorthogonal wavelets.

   Let $\{f_k\}$ be a set of functions in $L_2(\mathbb{R})$. If there are positive constants A and B such that

$$A \sum_k |c_k|^2 \leq \left\| \sum_k c_k f_k \right\|^2 \leq B \sum_k |c_k|^2, \tag{9.4.1}$$

for any $\{c_k\}$ in ℓ^2, then $\{f_k\}$ is called a **Riesz sequence**. In this case, A, B are called Riesz bounds. If $\{f_k\}$ satisfies the right-hand inequality of (9.4.1), then it is called a **Bessel sequence** with a bound B.

Definition 1 $\boxed{\text{\textbf{Riesz basis}}}$ *A set $\{f_k\}$ of functions in $L_2(\mathbb{R})$ is called a Riesz basis for $L_2(\mathbb{R})$ if it is a Riesz sequence and it is complete in $L_2(\mathbb{R})$.*

Clearly, if $\{f_k\}$ satisfies (9.4.1), then it is linearly independent (see Exercise 1). Thus, a Riesz basis for $L_2(\mathbb{R})$ is a basis for $L_2(\mathbb{R})$.

Recall that two sets $\{f_k\}$ and $\{\widetilde{f}_k\}$ of functions in $L_2(\mathbb{R})$ are said to be **biorthogonal to each other** or to form **biorthogonal systems** if

$$\langle f_j, \widetilde{f}_k \rangle = \delta_{j-k}.$$

For two biorthogonal systems $\{f_k\}$ and $\{\widetilde{f}_k\}$, to verify whether they are Riesz sequences, we only need to check whether they are Bessel sequences, as stated in the next theorem.

Theorem 1 $\boxed{\text{\textbf{Biorthogonal Bessel sequences are Riesz sequences}}}$ *Let $\{f_k\}$ and $\{g_k\}$ be two sets of functions in $L_2(\mathbb{R})$ which form biorthogonal systems of $L_2(\mathbb{R})$. Then $\{f_k\}$ and $\{g_k\}$ are Riesz sequences if and only if they are Bessel sequences.*

Proof Suppose

$$\left\| \sum_k c_k f_k \right\|^2 \le B \sum_k |c_k|^2, \ \left\| \sum_k c_k g_k \right\|^2 \le B' \sum_k |c_k|^2,$$

for any $\sum_k |c_k|^2 < \infty$. By the biorthogonality of $\{f_k\}$ and $\{g_k\}$, together with the Cauchy-Schwarz inequality, we have

$$\sum_k |c_k|^2 = \langle \sum_j c_j f_j, \sum_k c_k g_k \rangle \le \left\| \sum_j c_j f_j \right\| \left\| \sum_k c_k g_k \right\|$$

$$\le \sqrt{B} \left(\sum_k |c_k|^2 \right)^{\frac{1}{2}} \left\| \sum_k c_k g_k \right\|.$$

Thus

$$\frac{1}{B} \sum_k |c_k|^2 \le \left\| \sum_k c_k g_k \right\|^2.$$

Therefore, $\{g_k\}$ is a Riesz sequence with bounds B^{-1}, B'. Similarly, one can show that $\{f_k\}$ is a Riesz sequence with bounds B'^{-1}, B. ∎

For a function ϕ in $L_2(\mathbb{R})$, from the definition of stability of ϕ (see (8.2.17) on p.394), we know that ϕ is stable if $\{\phi(x-k) : k \in \mathbb{Z}\}$ is a Riesz sequence. Also, recall that two functions ϕ and $\widetilde{\phi}$ in $L_2(\mathbb{R})$ are said to be biorthogonal to each other if $\{\phi(x-k) : k \in \mathbb{Z}\}$ and $\{\widetilde{\phi}(x-k) : k \in \mathbb{Z}\}$ form biorthogonal systems of $L_2(\mathbb{R})$, that is,

$$\langle \phi(\cdot - j), \widetilde{\phi}(\cdot - k) \rangle = \delta_{j-k}, \ j, k \in \mathbb{Z}. \tag{9.4.2}$$

For $\phi \in L_2(\mathbb{R})$, it is shown in the proof of (i) in Theorem 4, on p.449, that $\{\phi(x-k) : k \in \mathbb{Z}\}$ is a Bessel sequence with bound B if and only if

$$G_\phi(\omega) \leq B, \ a.e. \ \omega \in \mathbb{R},$$

where $G_\phi(\omega)$ is the Gramian function of ϕ defined by (9.2.1) on p.445. In addition, if ϕ is compactly supported, then $G_\phi(\omega)$ is a trigonometric polynomial (see Remark 1 on p.447), and hence, it is upper bounded. Thus $\{\phi(x - k) : \ k \in \mathbb{Z}\}$ is a Bessel sequence. This observation, together with Theorem 1, leads to the following corollary.

Corollary 1 $\boxed{\textbf{Biorthogonality implies stability}}$ *If two compactly supported functions ϕ and $\widetilde{\phi}$ are biorthogonal to each other, then ϕ and $\widetilde{\phi}$ are stable.*

For a real-valued function $\psi \in L_2(\mathbb{R})$, denote, as before,

$$\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k), \quad j, k \in \mathbb{Z}.$$

We say that ψ is a **wavelet** provided that $\psi_{j,k}, \ j, k \in \mathbb{Z}$ form a Riesz basis of $L_2(\mathbb{R})$. If, in addition, two families of functions $\{\psi_{j,k}, \ j, k \in \mathbb{Z}\}$ and $\{\widetilde{\psi}_{j,k}, \ j, k \in \mathbb{Z}\}$ are biorthogonal to each other, namely,

$$\langle \psi_{j,k}, \widetilde{\psi}_{j',k'} \rangle = \delta_{j-j'}\delta_{k-k'}, \ \ j, k \in \mathbb{Z}, \tag{9.4.3}$$

then two wavelets ψ and $\widetilde{\psi}$ are called a pair of **biorthogonal wavelets**.

Since a Riesz basis of $L_2(\mathbb{R})$ is a basis for $L_2(\mathbb{R})$, we know that $\{\psi_{j,k}\}$ and $\{\widetilde{\psi}_{j,k}\}, \ j, k \in \mathbb{Z}$, are two bases for $L_2(\mathbb{R})$. Thus, any $f \in L_2(\mathbb{R})$ can be expanded by either $\{\psi_{j,k}\}$ or $\{\widetilde{\psi}_{j,k}\}, \ j, k \in \mathbb{Z}$, and the biorthogonality property gives the expansion coefficients as the inner products of f with $\widetilde{\psi}_{j,k}$ or with $\psi_{j,k}$ (see Exercise 2):

$$f(x) = \sum_{j,k \in \mathbb{Z}} \langle f, \widetilde{\psi}_{j,k} \rangle \psi_{j,k}(x) = \sum_{j,k \in \mathbb{Z}} \langle f, \psi_{j,k} \rangle \widetilde{\psi}_{j,k}(x). \tag{9.4.4}$$

In addition, the Riesz basis properties of $\widetilde{\psi}_{j,k}$ and $\psi_{j,k}$ imply that they satisfy the frame conditions (see Exercise 3)

$$\frac{1}{\widetilde{B}}\|f\|^2 \leq \sum_{j,k \in \mathbb{Z}} |\langle f, \psi_{j,k} \rangle|^2 \leq \frac{1}{\widetilde{A}}\|f\|^2, \tag{9.4.5}$$

$$\frac{1}{B}\|f\|^2 \leq \sum_{j,k \in \mathbb{Z}} |\langle f, \widetilde{\psi}_{j,k} \rangle|^2 \leq \frac{1}{A}\|f\|^2,$$

where A, B and $\widetilde{A}$, $\widetilde{B}$ are the Riesz basis bounds for $\{\psi_{j,k}\}$ and $\{\widetilde{\psi}_{j,k}\}$, respectively. The frame conditions assure the stability of the wavelet expansions.

By Theorem 1, to construct biorthogonal wavelets, we need to construct ψ and $\widetilde{\psi}$ such that:

(a) Each collection of $\{\psi_{j,k}\}, \ j, k \in \mathbb{Z}$, and $\{\widetilde{\psi}_{j,k}\}, \ j, k \in \mathbb{Z}$, is complete for $L_2(\mathbb{R})$;
(b) ψ and $\widetilde{\psi}$ satisfy (9.4.3);

(c) $\{\psi_{j,k}\}$, $j, k \in \mathbb{Z}$, and $\{\widetilde{\psi}_{j,k}\}$, $j, k \in \mathbb{Z}$, are Bessel sequences.

For a compactly supported ψ, it can be shown that under some mild conditions on ψ, $\{\psi_{j,k} : j, k \in \mathbb{Z}\}$ is a Bessel sequence.

Theorem 2 $\boxed{\textbf{Bessel sequence}}$ *Let $\psi \in L_2(\mathbb{R})$ be a compactly supported function. Suppose ψ satisfies*

(i) $\displaystyle\int_{-\infty}^{\infty} \psi(x)dx = 0$; *and*

(ii) *there is a $\gamma > 0$ such that*

$$\int_{-\infty}^{\infty} |\widehat{\psi}(\omega)|^2 (1 + \omega^2)^\gamma d\omega < \infty.$$

Then $\{\psi_{j,k} : j, k \in \mathbb{Z}\}$ is a Bessel sequence of $L_2(\mathbb{R})$.

Theorem 2 states that, for a compactly supported function ψ in $L_2(\mathbb{R})$, if it has a vanishing moment of order 1 and certain smoothness, then $\{\psi_{j,k} : j, k \in \mathbb{Z}\}$ is a Bessel sequence of $L_2(\mathbb{R})$. The biorthogonal wavelets to be constructed do satisfy these mild conditions.

Next, let us look at the MRA-based construction of biorthogonal wavelets. In this case we construct two MRAs $\{\mathbb{V}_j\}$ and $\{\widetilde{\mathbb{V}}_j\}$, which are biorthogonal to each other in the sense that their compactly supported scaling functions ϕ and $\widetilde{\phi}$ are biorthogonal to each other, namely, they satisfy (9.4.2). With

$$\mathbb{V}_j = \overline{\text{span}}\{\phi(2^j \cdot -k) : k \in \mathbb{Z}\}, \ \widetilde{\mathbb{V}}_j = \overline{\text{span}}\{\widetilde{\phi}(2^j \cdot -k) : k \in \mathbb{Z}\}, \qquad (9.4.6)$$

if we can construct compactly supported functions $\psi, \widetilde{\psi} \in L_2(\mathbb{R})$ such that

$$\langle \psi, \widetilde{\psi}(\cdot - k)\rangle = \delta_k, \ k \in \mathbb{Z}; \qquad\qquad (9.4.7)$$
$$\langle \phi, \widetilde{\psi}(\cdot - k)\rangle = 0, \ \langle \widetilde{\phi}, \psi(\cdot - k)\rangle = 0, \ k \in \mathbb{Z}, \qquad\qquad (9.4.8)$$

and such that $\mathbb{V}_1$ and $\widetilde{\mathbb{V}}_1$ are the direct sums of their subspaces $\mathbb{W}_0$ and $\mathbb{V}_0$, and $\widetilde{\mathbb{W}}_0$ and $\widetilde{\mathbb{V}}_0$, respectively, that is,

$$\mathbb{V}_1 = \mathbb{W}_0 \oplus \mathbb{V}_0, \ \widetilde{\mathbb{V}}_1 = \widetilde{\mathbb{W}}_0 \oplus \widetilde{\mathbb{V}}_0, \qquad\qquad (9.4.9)$$

where

$$\mathbb{W}_0 = \overline{\text{span}}\{\psi(\cdot - k) : k \in \mathbb{Z}\}, \ \widetilde{\mathbb{W}}_0 = \overline{\text{span}}\{\widetilde{\psi}(\cdot - k) : k \in \mathbb{Z}\},$$

then ψ and $\widetilde{\psi}$ form a pair of biorthogonal wavelets. A vector space $\mathbb{V}$ is said to be a **direct sum** of its two subspaces $\mathbb{U}$ and $\mathbb{W}$, which is denoted as

$$\mathbb{V} = \mathbb{U} \oplus \mathbb{W},$$

provided that each $\mathbf{v} \in \mathbb{V}$ can be written as $\mathbf{v} = \mathbf{u} + \mathbf{w}$, with $\mathbf{u} \in \mathbb{U}$, $\mathbf{w} \in \mathbb{W}$ and $\mathbb{U} \cap \mathbb{W} = \{\mathbf{0}\}$.

Theorem 3 $\boxed{\textbf{Biorth. wavelets derived from biorth. scaling functions}}$ *Suppose two compactly supported refinable functions $\phi, \widetilde{\phi} \in L_2(\mathbb{R})$ are biorthogonal to each other. If two compactly supported functions $\psi, \widetilde{\psi} \in L_2(\mathbb{R})$ satisfy* (9.4.7)–(9.4.9), *then ψ and $\widetilde{\psi}$ are a pair of biorthogonal wavelets, i.e. $\{\psi_{j,k}\}$ and $\{\widetilde{\psi}_{j,k}\}$, $j, k \in \mathbb{Z}$, are Riesz bases for $L_2(\mathbb{R})$, and satisfy* (9.4.3).

Proof   By Corollary 1, we know that both ϕ and $\widetilde{\phi}$ are stable, since they are biorthogonal to each other. Thus $\{\mathbb{V}_j\}$ and $\{\widetilde{\mathbb{V}}_j\}$, defined by (9.4.6), are two MRAs. Denote

$$\mathbb{W}_j = \overline{\text{span}}\{\psi(2^j \cdot -k) : k \in \mathbb{Z}\}, \;\; \widetilde{\mathbb{W}}_j = \overline{\text{span}}\{\widetilde{\psi}(2^j \cdot -k) : k \in \mathbb{Z}\}.$$

Then (9.4.9) implies

$$\mathbb{V}_{j+1} = \mathbb{W}_j \oplus \mathbb{V}_j, \;\; \widetilde{\mathbb{V}}_{j+1} = \widetilde{\mathbb{W}}_j \oplus \widetilde{\mathbb{V}}_j.$$

This, together with conditions $(2°)$ and $(3°)$ of MRA, implies that

$$L_2(\mathbb{R}) = \sum_{j \in \mathbb{Z}} \oplus \mathbb{W}_j, \;\; L_2(\mathbb{R}) = \sum_{j \in \mathbb{Z}} \oplus \widetilde{\mathbb{W}}_j. \tag{9.4.10}$$

The biorthogonality property (9.4.8) implies

$$\mathbb{W}_j \perp \widetilde{\mathbb{V}}_j, \;\; \widetilde{\mathbb{W}}_j \perp \mathbb{V}_j. \tag{9.4.11}$$

This and (9.4.7) lead to

$$\mathbb{W}_j \perp \widetilde{\mathbb{W}}_{j'}, \;\; j, j' \in \mathbb{Z}. \tag{9.4.12}$$

Indeed, if $j = j'$, (9.4.12) follows from (9.4.7). For $j \neq j'$, assuming $j' < j$, we have

$$\widetilde{\mathbb{W}}_{j'} \subset \widetilde{\mathbb{V}}_j.$$

This and $\widetilde{\mathbb{V}}_j \perp \mathbb{W}_j$ imply $\widetilde{\mathbb{W}}_{j'} \perp \mathbb{W}_j$, as desired.

Then from (9.4.10), we see that each of $\{\psi_{j,k} : j, k \in \mathbb{Z}\}$ and $\{\widetilde{\psi}_{j,k} : j, k \in \mathbb{Z}\}$ is complete in $L_2(\mathbb{R})$. Furthermore, (9.4.12) implies that $\{\psi_{j,k} : j, k \in \mathbb{Z}\}$ and $\{\widetilde{\psi}_{j,k} : j, k \in \mathbb{Z}\}$ are biorthogonal systems of $L_2(\mathbb{R})$ (see Exercise 4). Thus, conditions (a) and (b) listed on p.467 for ψ and $\widetilde{\psi}$ to be biorthogonal wavelets are fulfilled. For condition (c), we will find that biorthogonal compactly supported wavelets automatically satisfy the vanishing moment and smoothness conditions (i) and (ii) in Theorem 2. Therefore, ψ and $\widetilde{\psi}$ are biorthogonal wavelets. $\blacksquare$

Except for the biorthogonality conditions (9.4.2), (9.4.7) and (9.4.8) on $\phi, \widetilde{\phi}, \psi, \widetilde{\psi}$, and the refinement properties of $\phi, \widetilde{\phi}$, (9.4.9) also needs to be satisfied. Condition (9.4.9) is equivalent to

$$\psi(x) = \sum_k q_k \phi(2x - k), \ \ \widetilde{\psi}(x) = \sum_k \widetilde{q}_k \widetilde{\phi}(2x - k), \qquad (9.4.13)$$

$$\begin{cases} \phi(2x) = \sum_k c_k \phi(x - k) + \sum_k d_k \psi(x - k), \\ \widetilde{\phi}(2x) = \sum_k \widetilde{c}_k \widetilde{\phi}(x - k) + \sum_k \widetilde{d}_k \widetilde{\psi}(x - k), \end{cases} \qquad (9.4.14)$$

where $\{q_k\}$, $\{\widetilde{q}_k\}$ are finite sequences of real numbers, and c_k, d_k, $\widetilde{c}_k$, $\widetilde{d}_k$ are some real numbers such that the series in (9.4.14) converge in $L_2(\mathbb{R})$. The direct sum property in (9.4.9) also requires $\mathbb{V}_0 \cap \mathbb{W}_0 = \{0\}$, $\widetilde{\mathbb{V}}_0 \cap \widetilde{\mathbb{W}}_0 = \{0\}$, which can be easily obtained from the biorthogonality of ϕ, $\widetilde{\phi}$, ψ, $\widetilde{\psi}$ (see Exercise 5).

To construct biorthogonal wavelets, we start with the masks $\{p_k\}$ and $\{\widetilde{p}_k\}$ for the scaling functions ϕ and $\widetilde{\phi}$. Next, we derive the sequences $\{q_k\}$ and $\{\widetilde{q}_k\}$ from $\{p_k\}$ and $\{\widetilde{p}_k\}$ such that ψ and $\widetilde{\psi}$ are given by (9.4.13). Since we focus on compactly supported wavelets, $\{p_k\}$, $\{\widetilde{p}_k\}$, $\{q_k\}$ and $\{\widetilde{q}_k\}$ are finite sequences. Let $p(\omega)$, $\widetilde{p}(\omega)$, $q(\omega)$ and $\widetilde{q}(\omega)$ be the corresponding two-scale symbols. Then (9.4.13) can be written as

$$\widehat{\psi}(\omega) = q\left(\frac{\omega}{2}\right)\widehat{\phi}\left(\frac{\omega}{2}\right), \ \ \widehat{\widetilde{\psi}}(\omega) = \widetilde{q}\left(\frac{\omega}{2}\right)\widehat{\widetilde{\phi}}\left(\frac{\omega}{2}\right). \qquad (9.4.15)$$

For ϕ, $\widetilde{\phi} \in L_2(\mathbb{R})$, let $G_{\widetilde{\phi},\phi}(\omega)$ be the 2π-periodic function defined by (9.2.2) in Sect. 9.2, on p.445. Recall that, from (9.2.4) in Theorem 1 of Sect. 9.2, ϕ and $\widetilde{\phi}$ are biorthogonal if and only if

$$G_{\widetilde{\phi},\phi}(\omega) = 1, a.e. \ \omega \in \mathbb{R}.$$

Theorem 4 $\boxed{\text{Biorthogonality implies perfect reconstruction}}$ *Let ϕ, $\widetilde{\phi} \in L_2(\mathbb{R})$ be refinable functions with refinement masks $\{p_k\}$, $\{\widetilde{p}_k\}$. Let ψ and $\widetilde{\psi}$ be the functions defined by (9.4.13). Suppose ϕ and $\widetilde{\phi}$ are biorthogonal to each other, i.e. they satisfy (9.4.2). Then*

$$\widetilde{p}(\omega)\overline{p(\omega)} + \widetilde{p}(\omega + \pi)\overline{p(\omega + \pi)} = 1. \qquad (9.4.16)$$

Furthermore, (9.4.7) and (9.4.8) are satisfied if and only if

$$\widetilde{q}(\omega)\overline{q(\omega)} + \widetilde{q}(\omega + \pi)\overline{q(\omega + \pi)} = 1, \qquad (9.4.17)$$

$$\widetilde{q}(\omega)\overline{p(\omega)} + \widetilde{q}(\omega + \pi)\overline{p(\omega + \pi)} = 0, \qquad (9.4.18)$$

$$\widetilde{p}(\omega)\overline{q(\omega)} + \widetilde{p}(\omega + \pi)\overline{q(\omega + \pi)} = 0. \qquad (9.4.19)$$

Proof To show (9.4.16), one can follow the proof of (9.3.10) on p.454 by using

$$G_{\widetilde{\phi},\phi}(2\omega) = \widetilde{p}(\omega)\overline{p(\omega)}G_{\widetilde{\phi},\phi}(\omega) + \widetilde{p}(\omega + \pi)\overline{p(\omega + \pi)}G_{\widetilde{\phi},\phi}(\omega + \pi).$$

Thus, if ϕ and $\widetilde{\phi}$ are biorthogonal to each other, or equivalently, $G_{\widetilde{\phi},\phi}(\omega) = 1$, then the above equation implies that $p(\omega)$ and $\widetilde{p}(\omega)$ satisfy (9.4.16).

Similarly, by

$$G_{\widetilde{\psi},\psi}(2\omega) = \widetilde{q}(\omega)\overline{q(\omega)}G_{\widetilde{\phi},\phi}(\omega) + \widetilde{q}(\omega+\pi)\overline{q(\omega+\pi)}G_{\widetilde{\phi},\phi}(\omega+\pi),$$

$$G_{\widetilde{\psi},\phi}(2\omega) = \widetilde{q}(\omega)\overline{p(\omega)}G_{\widetilde{\phi},\phi}(\omega) + \widetilde{q}(\omega+\pi)\overline{p(\omega+\pi)}G_{\widetilde{\phi},\phi}(\omega+\pi),$$

$$G_{\widetilde{\phi},\psi}(2\omega) = \widetilde{p}(\omega)\overline{q(\omega)}G_{\widetilde{\phi},\phi}(\omega) + \widetilde{p}(\omega+\pi)\overline{q(\omega+\pi)}G_{\widetilde{\phi},\phi}(\omega+\pi),$$

we see that, with the assumption $G_{\widetilde{\phi},\phi}(\omega) = 1$, the biorthogonality conditions (9.4.7) and (9.4.8), which are equivalent to

$$G_{\widetilde{\psi},\psi}(\omega) = 1, \ \ G_{\widetilde{\psi},\phi}(\omega) = 0, \ \ G_{\widetilde{\phi},\psi}(\omega) = 0,$$

are satisfied if and only if (9.4.17)–(9.4.19) hold.   ∎

Recall that if lowpass filters (refinement masks) p, $\widetilde{p}$ satisfy (9.4.16), we say that they are biorthogonal; and if $\widetilde{p}, \widetilde{q}, p, q$ satisfy (9.4.16)–(9.4.19), then we say they form a biorthogonal filter bank or a perfect reconstruction (PR) filter bank. As discussed in Sect. 8.2, for a given pair of biorthogonal FIR lowpass filters p and $\widetilde{p}$, we may choose the FIR highpass filters q and $\widetilde{q}$ to be

$$q(\omega) = -\overline{\widetilde{p}(\omega+\pi)}e^{-i(2s-1)\omega}, \ \ \widetilde{q}(\omega) = -\overline{p(\omega+\pi)}e^{-i(2s-1)\omega},$$

or equivalently,

$$q_k = (-1)^k \widetilde{p}_{2s-1-k}, \ \ \widetilde{q}_k = (-1)^k \widetilde{p}_{2s-1-k}, \ \ k \in \mathbb{Z}, \qquad (9.4.20)$$

where s is an integer such that $\widetilde{p}, \widetilde{q}, p, q$ form a biorthogonal (PR) filter bank.

Up to now, we have seen that the key to constructing biorthogonal wavelets is to construct finite sequences $\{p_k\}$ and $\{\widetilde{p}_k\}$ such that the associated refinable functions ϕ and $\widetilde{\phi}$ are biorthogonal. Next, we give a characterization on the biorthogonality of ϕ and $\widetilde{\phi}$, the proof of which will be provided in Chap. 10, on p.528.

Theorem 5 ⏐**Biorthogonality of refinable functions**⏐ *Let ϕ and $\widetilde{\phi}$ be compactly supported refinable functions associated with some trigonometric polynomials $p(\omega)$ and $\widetilde{p}(\omega)$ satisfying $p(0) = 1$, $\widetilde{p}(0) = 1$. Then ϕ and $\widetilde{\phi}$ are in $L_2(\mathbb{R})$ and biorthogonal to each other if and only if*

 (i) *$p(\omega)$ and $\widetilde{p}(\omega)$ satisfy (9.4.16);*
 (ii) *$p(\pi) = 0$ and $\widetilde{p}(\pi) = 0$; and*
(iii) *T_p and $T_{\widetilde{p}}$ satisfy Condition E.*

Theorem 6 *Let $\phi, \widetilde{\phi}$ be compactly supported refinable functions associated with $p(\omega), \widetilde{p}(\omega)$ satisfying (i)–(iii) in Theorem 5. Let $\psi, \widetilde{\psi}$ be the functions given by (9.4.13) with $\{q_k\}, \{\widetilde{q}_k\}$ defined by (9.4.20). Then ψ and $\widetilde{\psi}$ are a pair of compactly supported biorthogonal wavelets.*

Proof By Theorem 5, we know that ϕ and $\widetilde{\phi}$ are in $L_2(\mathbb{R})$ and biorthogonal to each other (which implies that they are stable). Thus, by Theorem 4, (9.4.7) and (9.4.8) are satisfied, since $q(\omega)$ and $\widetilde{q}(\omega)$ satisfy (9.4.17)–(9.4.19). Therefore, by Theorem 3, ψ and $\widetilde{\psi}$ are a pair of compactly supported biorthogonal wavelets if (9.4.14) is satisfied. Next, let us verify (9.4.14).

Indeed, from (9.4.16)–(9.4.19), we have

$$M_{\widetilde{p},\widetilde{q}}(\omega)\,\overline{M_{p,q}(\omega)}^T = I_2, \ \omega \in \mathbb{R},$$

where $M_{\widetilde{p},\widetilde{q}}$ and $M_{p,q}$ are modulation matrices of $\widetilde{p},\widetilde{q}$ and p,q, respectively, defined by (9.3.12) on p.455. Thus $\overline{M_{p,q}(\omega)}^T$ is the inverse matrix of $M_{\widetilde{p},\widetilde{q}}(\omega)$, and we have

$$\overline{M_{p,q}(\omega)}^T M_{\widetilde{p},\widetilde{q}}(\omega) = I_2, \ \omega \in \mathbb{R}.$$

In particular, we have

$$\overline{p(\omega)}\widetilde{p}(\omega) + \overline{q(\omega)}\widetilde{q}(\omega) = 1, \ \overline{p(\omega)}\widetilde{p}(\omega+\pi) + \overline{q(\omega)}\widetilde{q}(\omega+\pi) = 0,$$

or

$$\overline{\widetilde{p}(\omega)}p(\omega) + \overline{\widetilde{q}(\omega)}q(\omega) = 1, \ \overline{\widetilde{p}(\omega+\pi)}p(\omega) + \overline{\widetilde{q}(\omega+\pi)}q(\omega) = 0. \quad (9.4.21)$$

Therefore,

$$\left(\overline{\widetilde{p}(\omega)} + \overline{\widetilde{p}(\omega+\pi)}\right)p(\omega) + \left(\overline{\widetilde{q}(\omega)} + \overline{\widetilde{q}(\omega+\pi)}\right)q(\omega) = 1.$$

Hence,

$$\begin{aligned}
\widehat{\phi}\left(\frac{\omega}{2}\right) &= \left(\overline{\widetilde{p}\left(\frac{\omega}{2}\right)} + \overline{\widetilde{p}\left(\frac{\omega}{2}+\pi\right)}\right)p\left(\frac{\omega}{2}\right)\widehat{\phi}\left(\frac{\omega}{2}\right) + \left(\overline{\widetilde{q}\left(\frac{\omega}{2}\right)} + \overline{\widetilde{q}\left(\frac{\omega}{2}+\pi\right)}\right)q\left(\frac{\omega}{2}\right)\widehat{\phi}\left(\frac{\omega}{2}\right) \\
&= \frac{1}{2}\sum_k \left(1+(-1)^k\right)\overline{\widetilde{p}_k}e^{ik\frac{\omega}{2}}\,\widehat{\phi}(\omega) + \frac{1}{2}\sum_k \left(1+(-1)^k\right)\overline{\widetilde{q}_k}e^{ik\frac{\omega}{2}}\,\widehat{\psi}(\omega) \\
&= \sum_n \widetilde{p}_{2n}e^{in\omega}\,\widehat{\phi}(\omega) + \sum_n \widetilde{q}_{2n}e^{in\omega}\,\widehat{\psi}(\omega),
\end{aligned}$$

where we use the assumption that $\widetilde{p}_k, \widetilde{q}_k$ are real in the last equality. Thus,

$$\phi(2x) = \frac{1}{2}\sum_n \widetilde{p}_{-2n}\phi(x-n) + \frac{1}{2}\sum_n \widetilde{q}_{-2n}\psi(x-n).$$

Similarly, we have

$$\widetilde{\phi}(2x) = \frac{1}{2}\sum_n p_{-2n}\widetilde{\phi}(x-n) + \frac{1}{2}\sum_n q_{-2n}\widetilde{\psi}(x-n).$$

This shows that (9.4.14) holds, as desired.   ■

Next, we consider the symmetric or antisymmetric biorthogonal wavelets. Let $N \in \mathbb{Z}$. We say a sequence $\{h_k\}$ is symmetric around $\frac{N}{2}$ provided that $h_{N-k} = h_k, k \in \mathbb{Z}$, and we say it is antisymmetric if $h_{N-k} = -h_k, k \in \mathbb{Z}$. Clearly, for a (finite) sequence $\{h_k\}$, it is symmetric or antisymmetric around $\frac{N}{2}$ if and only if the two-scale symbol $h(\omega)$ satisfies

$$h(-\omega) = e^{iN\omega}h(\omega) \text{ or } h(-\omega) = -e^{iN\omega}h(\omega)$$

(see Exercise 7).

Theorem 7 $\boxed{\textbf{Symmetry}}$ *Let ϕ be the normalized refinable function associated with a finite sequence $\{p_k\}$. If $\{p_k\}$ is symmetric around $\frac{N}{2}$ for some integer N, then ϕ is also symmetric around $\frac{N}{2}$, i.e.*

$$\phi(x) = \phi(N - x).$$

If, in addition, $\{q_k\}$ is symmetric/antisymmetric around $\frac{N'}{2}$, then ψ, defined by

$$\psi(x) = \sum_k q_k \phi(2x - k),$$

is symmetric/antisymmetric around $\frac{1}{4}(N + N')$.

Proof We start by noting that

$$\widehat{\phi}(-\omega) = \prod_{j=1}^{\infty} p\left(-\frac{\omega}{2^j}\right) = \prod_{j=1}^{\infty} e^{iN\frac{\omega}{2^j}} p\left(\frac{\omega}{2^j}\right) = e^{iN\omega}\widehat{\phi}(\omega).$$

Thus $\phi(x) = \phi(N - x)$, as desired.

From $\widehat{\psi}(\omega) = q\left(\frac{\omega}{2}\right)\widehat{\phi}\left(\frac{\omega}{2}\right)$ and $q(-\omega) = \pm e^{iN'\omega}q(\omega)$, we have

$$\widehat{\psi}(-\omega) = q(-\frac{\omega}{2})\widehat{\phi}(-\frac{\omega}{2}) = \pm e^{iN'\omega/2}q\left(\frac{\omega}{2}\right)e^{iN\omega/2}\widehat{\phi}\left(\frac{\omega}{2}\right) = \pm e^{i\frac{(N+N')}{2}\omega}\widehat{\psi}(\omega).$$

Therefore $\psi(x) = \pm\psi\left(\frac{N+N'}{2} - x\right)$, that is, ψ is symmetric or antisymmetric around $\frac{1}{4}(N + N')$ (depending on the $+$ or $-$ sign). ∎

One can show that if $\{p_k\}$ and $\{\widetilde{p}_k\}$ are symmetric, then $\{q_k\}$ and $\{\widetilde{q}_k\}$, defined by (9.4.20), are symmetric/antisymmetric (see Exercise 8).

$\boxed{\textbf{Sum-rules imply vanishing moments}}$ Except for symmetry, we also require that the constructed biorthogonal wavelets have the vanishing moment property. Suppose $p(\omega), q(\omega), \widetilde{p}(\omega), \widetilde{q}(\omega)$ form a biorthogonal FIR filter bank. Then one can show that if $p(\omega)$ has sum-rule order L, then $\widetilde{q}(\omega)$ satisfies

$$\frac{d^\ell}{d\omega^\ell}\widetilde{q}(0) = 0, \text{ for } 0 \leq \ell \leq L - 1 \tag{9.4.22}$$

(see Exercise 10). Hence, $\widetilde{\psi}$ has vanishing moments of order L (refer to Theorem 5 on p.457). Similarly, if $\widetilde{p}(\omega)$ has sum-rule order $\widetilde{L}$, then ψ has vanishing moments of order $\widetilde{L}$. Hence, to construct biorthogonal wavelets with vanishing moments, we only need to construct lowpass filters with sum-rule orders. ∎

Next, we construct symmetric lowpass filters $\{p_k\}$ and $\{\widetilde{p}_k\}$ with certain sum-rule orders.

Construction of symmetric biorthogonal wavelets

Suppose $p(\omega)$ and $\widetilde{p}(\omega)$ are given by

$$p(\omega) = \cos^\ell \frac{\omega}{2} \, p_0(\omega), \, \widetilde{p}(\omega) = \cos^{\widetilde{\ell}} \frac{\omega}{2} \, \widetilde{p}_0(\omega),$$

where $p_0(\omega)$ and $\widetilde{p}_0(\omega)$ are some trigonometric polynomials, and $\ell, \widetilde{\ell}$ are positive integers such that $\ell + \widetilde{\ell}$ is even, that is,

$$\ell + \widetilde{\ell} = 2L$$

for a positive integer L. Because of symmetry, $p_0(\omega)\overline{\widetilde{p}_0(\omega)}$ is even, and hence it can be written as a function of $\cos\omega$. With $\cos\omega = 1 - 2\sin^2 \frac{\omega}{2}$, $p_0(\omega)\overline{\widetilde{p}_0(\omega)}$ can be expressed as

$$p_0(\omega)\overline{\widetilde{p}_0(\omega)} = P(\sin^2 \frac{\omega}{2}),$$

where P is a polynomial. Thus the biorthogonality condition (9.4.16) on $p(\omega)$ and $\widetilde{p}(\omega)$ is given by

$$\cos^{2L} \frac{\omega}{2} \, P(\sin^2 \frac{\omega}{2}) + \sin^{2L} \frac{\omega}{2} P(\cos^2 \frac{\omega}{2}) = 1, \, \omega \in \mathbb{R},$$

or equivalently, with $y = \sin^2 \frac{\omega}{2}$, $P(y)$ satisfies (9.3.18) on p.458. We will choose $P(y) = P_{L-1}(y)$, where P_{L-1} is the polynomial of degree $L - 1$ defined by (9.3.19) on p.458. Then, we factorize $\cos^{2L} \frac{\omega}{2} P_{L-1}(\sin^2 \frac{\omega}{2})$ into $p(\omega)\overline{\widetilde{p}(\omega)}$, to obtain biorthogonal filters p and $\widetilde{p}$. ∎

Next, we study the construction of the most commonly used symmetric biorthogonal filters $p(\omega)$ and $\widetilde{p}(\omega)$, and in particular, the so-called 5/3-tap and 9/7-tap filters that have been adopted by the Digital Image Compression Standard, JPEG 2000, for lossless and lossy compressions, respectively. The first step is to factorize $\cos^{2L} \frac{\omega}{2} P_{L-1}(\sin^2 \frac{\omega}{2})$ into $p(\omega)\overline{\widetilde{p}(\omega)}$. We will again use the notation

$$z = e^{-i\omega},$$

and rely on the fact that

$$\cos^2\frac{\omega}{2} = \frac{1}{4}(1+z)(1+\frac{1}{z}), \quad \sin^2\frac{\omega}{2} = \frac{1}{4}(2-z-\frac{1}{z}).$$

For $L = 1$, since $P_0(y) = 1$, we have

$$p(\omega)\overline{\tilde{p}(\omega)} = \cos^2\frac{\omega}{2} = \frac{1}{4}(1+z)(1+\frac{1}{z}).$$

Thus, the only possible choice for $p(\omega)$ and $\tilde{p}(\omega)$ is

$$p(\omega) = \tilde{p}(\omega) = \frac{1}{2}(1+z).$$

In this case we obtain the Haar wavelet. In the next three examples, we consider the cases $L = 2, 3, 4$.

Example 1 $\boxed{L = 2}$ For $L = 2$, with $P_1(y) = 1 + 2y$, we have

$$p(\omega)\overline{\tilde{p}(\omega)} = \cos^4\frac{\omega}{2}\left(1 + 2\sin^2\frac{\omega}{2}\right) = \frac{1}{32}(1+z)^2(1+\frac{1}{z})^2(4-z-\frac{1}{z}).$$

We may choose

$$p(\omega) = \frac{1}{2}(1+z),$$

$$\tilde{p}(\omega) = \frac{1}{16}(1+\frac{1}{z})(1+z)^2(4-z-\frac{1}{z})$$

$$= -\frac{1}{16}\frac{1}{z^2} + \frac{1}{16}\frac{1}{z} + \frac{1}{2} + \frac{1}{2}z + \frac{1}{16}z^2 - \frac{1}{16}z^3,$$

or we may choose the lowpass filters, called

$$\boxed{5/3\text{-}\textbf{tap biorthogonal lowpass filters}:}$$

$$p(\omega) = \frac{1}{4}(1+z)^2, \tag{9.4.23}$$

$$\tilde{p}(\omega) = \frac{1}{8}(1+z)^2(4-z-\frac{1}{z})$$

$$= -\frac{1}{8}\frac{1}{z} + \frac{1}{4} + \frac{3}{4}z + \frac{1}{4}z^2 - \frac{1}{8}z^3.$$

The highpass filters $q(\omega)$ and $\tilde{q}(\omega)$ corresponding to $p(\omega)$, $\tilde{p}(\omega)$ in (9.4.23) can be obtained by applying (9.4.20) for an integer s. In the following, we provide $q(\omega)$ and $\tilde{q}(\omega)$ with $s = 1$ to obtain the highpass filters, called

$$\boxed{5/3\text{-}\textbf{tap biorthogonal highpass filters}}:$$

$$q(\omega) = -\frac{1}{8}(z^2 + \frac{1}{z^2}) - \frac{1}{4}(z + \frac{1}{z}) + \frac{3}{4},$$

$$\widetilde{q}(\omega) = -\frac{1}{4}z(1 - \frac{1}{z})^2 = \frac{1}{2} - \frac{1}{4}(z + \frac{1}{z}). \qquad (9.4.24)$$

See Fig. 8.3 in Sect. 8.3, on p.403, for the graphs of the corresponding scaling functions and biorthogonal wavelets.   ■

Example 2 $\boxed{L = 3}$ For $L = 3$, we have

$$p(\omega)\overline{\widetilde{p}(\omega)} = \cos^6 \frac{\omega}{2}\left(1 + 3\sin^2 \frac{\omega}{2} + 6\sin^4 \frac{\omega}{2}\right)$$

$$= \frac{1}{256}(1 + z)^3(1 + \frac{1}{z})^3\left(19 - 9(z + \frac{1}{z}) + \frac{3}{2}(z^2 + \frac{1}{z^2})\right).$$

We may choose

$$p(\omega) = \frac{1}{4}(1 + z)^2,$$

$$\widetilde{p}(\omega) = \frac{1}{64}(1 + z)^4 \frac{1}{z}\left(19 - 9(z + \frac{1}{z}) + \frac{3}{2}(z^2 + \frac{1}{z^2})\right)$$

$$= \frac{3}{128}(z^5 + \frac{1}{z^3}) - \frac{3}{64}(z^4 + \frac{1}{z^2}) - \frac{1}{8}(z^3 + \frac{1}{z}) + \frac{19}{64}(z^2 + 1) + \frac{45}{64}z,$$

or

$$p(\omega) = \frac{1}{8}(1 + z)^3, \qquad (9.4.25)$$

$$\widetilde{p}(\omega) = \frac{1}{32}(1 + z)^3\left(19 - 9(z + \frac{1}{z}) + \frac{3}{2}(z^2 + \frac{1}{z^2})\right)$$

$$= \frac{3}{64}(z^5 + \frac{1}{z^2}) - \frac{9}{64}(z^4 + \frac{1}{z}) - \frac{7}{64}(z^3 + 1) + \frac{45}{64}(z^2 + z),$$

or

$$p(\omega) = \frac{1}{16}(1 + z)^4,$$

$$\widetilde{p}(\omega) = \frac{1}{16}z(1 + z)^2\left(19 - 9(z + \frac{1}{z}) + \frac{3}{2}(z^2 + \frac{1}{z^2})\right) \qquad (9.4.26)$$

$$= \frac{3}{32}(z^5 + \frac{1}{z}) - \frac{3}{8}(z^4 + 1) + \frac{5}{32}(z^3 + z) + \frac{5}{4}z^2.$$

Observe that the scaling function ϕ associated with p in (9.4.25) is the quadratic B-spline, while that associated with (9.4.26) is the cubic B-spline. However, $\widetilde{\phi}$ and ϕ associated with $\widetilde{p}$ and p in (9.4.26) are not biorthogonal to each other, as shown in Chap. 10, on p.530. ∎

Example 3 $\boxed{L = 4}$ For $L = 4$, we have

$$
\begin{aligned}
p(\omega)\overline{\widetilde{p}(\omega)} &= \cos^8 \frac{\omega}{2}\left(1 + 4\sin^2 \frac{\omega}{2} + 10\sin^4 \frac{\omega}{2} + 20\sin^6 \frac{\omega}{2}\right) \\
&= \frac{1}{4^4}(1 + z)^4(1 + \frac{1}{z})^4 Q_3(z),
\end{aligned}
$$

where

$$
Q_3(z) = 13 - \frac{131}{16}(z + \frac{1}{z}) + \frac{5}{2}(z^2 + \frac{1}{z^2}) - \frac{5}{16}(z^3 + \frac{1}{z^3}).
$$

We may choose

$$
p(\omega) = \frac{1}{4}(1 + z)^2,
$$

$$
\widetilde{p}(\omega) = \frac{1}{64}(1 + \frac{1}{z})^2(1 + z)^4 Q_3(z),
$$

or

$$
p(\omega) = \frac{1}{8}(1 + z)^3,
$$

$$
\widetilde{p}(\omega) = \frac{1}{32}(1 + \frac{1}{z})(1 + z)^4 Q_3(z).
$$

Recall that with $y = \frac{1}{4}(2 - z - \frac{1}{z})$, $Q_3(z) = P_3(y)$, and $P_3(y)$ can be factorized as in (9.3.22) on p.462. Thus, we have

$$
Q_3(z) = -\frac{5}{16}\left(z + \frac{1}{z} - (2 - 4\theta)\right)\left(z^2 + \frac{1}{z^2} - (6 + 4\theta)(z + \frac{1}{z}) + 10 + 8\theta - \frac{4}{5\theta}\right),
$$

where θ is given by (9.3.23) on p.462; and we may choose $\widetilde{p}(\omega)$ and $p(\omega)$ to obtain the lowpass filters, called:

9/7-tap biorthogonal lowpass filters :

$$
p(\omega) = \frac{1}{2^4}(1 + z)^4 \frac{1}{4\theta}\left(z + \frac{1}{z} - (2 - 4\theta)\right),
$$

$$
\widetilde{p}(\omega) = \frac{1}{2^4}(1 + z)^4(-\frac{5}{16})4\theta\left(z^2 + \frac{1}{z^2} - (6 + 4\theta)(z + \frac{1}{z}) + 10 + 8\theta - \frac{4}{5\theta}\right).
$$

The numbers of nonzero coefficients $\widetilde{p}_k$ and p_k are 9 and 7, respectively. These two biorthogonal filters are called the 9/7-tap biorthogonal filters. The nonzero p_k, $\widetilde{p}_k$ are given by

$$p_{-1} = p_5 = -0.0912717631142501, \quad p_0 = p_4 = -0.0575435262285002,$$
$$p_1 = p_3 = 0.5912717631142501, \quad p_2 = 1.1150870524570004;$$
$$\widetilde{p}_{-2} = \widetilde{p}_6 = 0.0534975148216202, \quad \widetilde{p}_{-1} = \widetilde{p}_5 = -0.0337282368857499,$$
$$\widetilde{p}_0 = \widetilde{p}_4 = -0.1564465330579805, \quad \widetilde{p}_1 = \widetilde{p}_3 = 0.5337282368857499,$$
$$\widetilde{p}_2 = 1.2058980364727207.$$

The corresponding highpass filters $q(\omega)$ and $\widetilde{q}(\omega)$ can be obtained from (9.4.20) for some suitable integer s. In the following, we provide the nonzero q_k and $\widetilde{q}_k$ with $s = 3$ for the highpass filters, called:

9/7-**tap biorthogonal highpass filters** :
$$q_{-1} = q_7 = -0.0534975148216202, \quad q_0 = q_6 = -0.0337282368857499,$$
$$q_1 = q_5 = 0.1564465330579805, \quad q_2 = q_4 = 0.5337282368857499,$$
$$q_3 = -1.2058980364727207;$$
$$\widetilde{q}_0 = \widetilde{q}_6 = -0.0912717631142501, \; \widetilde{q}_1 = \widetilde{q}_5 = 0.0575435262285002,$$
$$\widetilde{q}_2 = \widetilde{q}_4 = 0.5912717631142501, \; \widetilde{q}_3 = -1.1150870524570004.$$

See Fig. 9.3 for the graphs of the corresponding scaling functions and biorthogonal wavelets. ∎

Exercises

Exercise 1   Show that if $\{f_1, f_2, \ldots\}$ in $L_2(\mathbb{R})$ satisfies (9.4.1), then it is linearly independent.

Exercise 2   Suppose ψ and $\widetilde{\psi}$ are a pair of biorthogonal wavelets in $L_2(\mathbb{R})$. Show that, for any $f \in L_2(\mathbb{R})$, (9.4.4) is satisfied.

Exercise 3   Suppose ψ and $\widetilde{\psi}$ are a pair of biorthogonal wavelets in $L_2(\mathbb{R})$. Show that $\{\psi_{j,k}\}$ and $\{\widetilde{\psi}_{j,k}\}$, $j, k \in \mathbb{Z}$, satisfy the frame condition in (9.4.5).

Exercise 4   Show that the condition in (9.4.12) implies that $\{\psi_{j,k} : j, k \in \mathbb{Z}\}$ and $\{\widetilde{\psi}_{j,k} : j, k \in \mathbb{Z}\}$ form biorthogonal systems of $L_2(\mathbb{R})$.

Exercise 5   Suppose $\phi, \widetilde{\phi}, \psi, \widetilde{\psi} \in L_2(\mathbb{R})$ satisfy (9.4.7), (9.4.2) and (9.4.8). Let $\mathbb{V}_0, \widetilde{\mathbb{V}}_0, \mathbb{W}_0$ and $\widetilde{\mathbb{W}}_0$ be the spaces spanned by the integer translates of $\phi, \widetilde{\phi}, \psi$ and $\widetilde{\psi}$, respectively. Show that $\mathbb{V}_0 \cap \mathbb{W}_0 = \{0\}$, $\widetilde{\mathbb{V}}_0 \cap \widetilde{\mathbb{W}}_0 = \{0\}$.
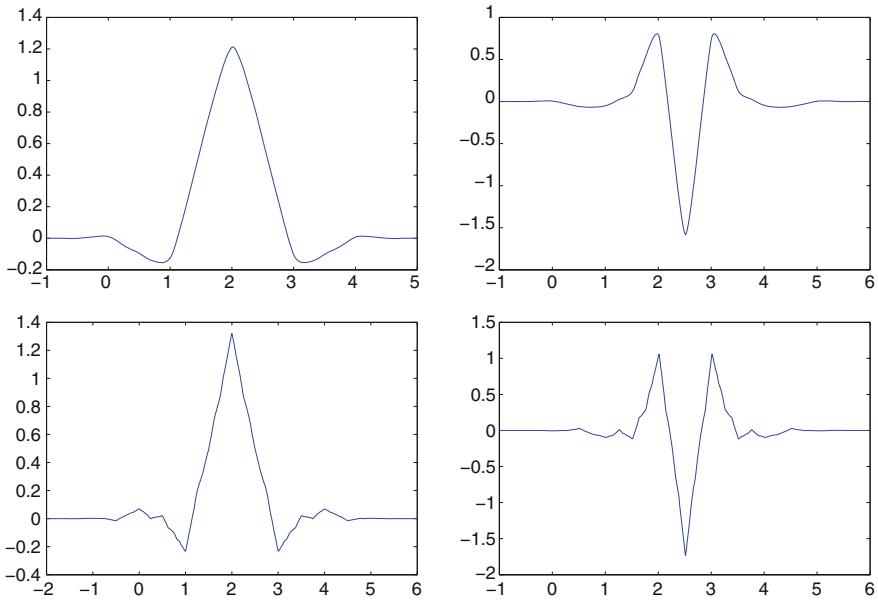
Fig. 9.3. *Top* biorthogonal 9/7 scaling function ϕ (on *left*) and wavelet ψ (on *right*); *Bottom* scaling function $\widetilde{\phi}$ (on *left*) and wavelet $\widetilde{\psi}$ (on *right*)

Exercise 6 Suppose that the FIR filters $p(\omega)$, $q(\omega)$, $\widetilde{p}(\omega)$ and $\widetilde{q}(\omega)$ are biorthogonal, that is, they satisfy (9.4.16)–(9.4.19). Show that, for any $n, m \in \mathbb{Z}$, $p(\omega)e^{-in\omega}$, $q(\omega)e^{-i(n+2m)\omega}$, $\widetilde{p}(\omega)e^{-in\omega}$, $\widetilde{q}(\omega)e^{-i(n+2m)\omega}$ are also biorthogonal.

Exercise 7 Show that a finite sequence $\{h_k\}$ is symmetric/antisymmetric around $\frac{N}{2}$ if and only if $h(-\omega) = \pm e^{iN\omega}h(\omega)$.

Exercise 8 Show that if $\{p_k\}$ and $\{\widetilde{p}_k\}$ are symmetric, then $\{q_k\}$ and $\{\widetilde{q}_k\}$, defined by (9.4.20), are symmetric/antisymmetric.

Exercise 9 Let $p(\omega)$, $q(\omega)$, $\widetilde{p}(\omega)$ and $\widetilde{q}(\omega)$ be the 5/3-tap biorthogonal filters given by (9.4.23)–(9.4.24). Verify directly that they satisfy (9.4.16)–(9.4.19).

Exercise 10 Suppose that the FIR filters $p(\omega)$, $\widetilde{q}(\omega)$ satisfy (9.4.18). Show that if $p(\omega)$ has sum-rule order L, then $\widetilde{q}(\omega)$ satisfies (9.4.22).
    Hint: Differentiate both sides of (9.4.18) with respect to ω and set $\omega = 0$.

9.5 Lifting Schemes

Recall that, for a filter or any sequence $\mathbf{g} = \{g_k\}$, the symbol $G(z)$ is used to denote its z-transform, defined by

$$G(z) = \sum_k g_k z^k.$$

Hence, if **g** is an FIR filter, then $G(z)$ is a Laurent polynomial in z; that is,

$$G(z) = \sum_{k=k_1}^{k_2} g_k z^k,$$

where k_1 and k_2 are integers, with $k_1 \leq k_2$, $g_{k_1} \neq 0$, $g_{k_2} \neq 0$ (k_1 could be a negative integer). We will call the degree of the Laurent polynomial G by the length of G,

$$\boxed{\text{length}(G) = k_2 - k_1},$$

since it is length of the filter **g** minus 1.

Also recall that if a sequence $h = \{h_k\}$ is a lowpass filter (refinement mask) for a refinable function or a highpass filter for a wavelet, the notation $h(\omega)$ is used to denote the two-scale symbol of h, defined by

$$h(\omega) = \frac{1}{2} \sum_k h_k e^{-ik\omega}.$$

For an FIR filter bank $\widetilde{p} = \{\widetilde{p}_k\}, \widetilde{q} = \{\widetilde{q}_k\}, p = \{p_k\}, q = \{q_k\}$, denote

$$\widetilde{\ell}_k = \frac{1}{2}\, \widetilde{p}_k, \ \widetilde{h}_k = \frac{1}{2}\, \widetilde{q}_k, \ \ell_k = p_k, \ h_k = q_k.$$

Let $\widetilde{L}(z), \widetilde{H}(z), L(z)$ and $H(z)$ be the z-transforms of $\widetilde{\ell}, \widetilde{h}, \ell$ and h, respectively. With $z = e^{-i\omega}$, we have

$$\widetilde{p}(\omega) = \widetilde{L}(z), \ \widetilde{q}(\omega) = \widetilde{H}(z),$$
$$p(\omega) = \frac{1}{2}L(z), \ q(\omega) = \frac{1}{2}H(z).$$

Recall that the discrete wavelet transform (DWT) (8.3.4) with $\widetilde{p}, \widetilde{q}$ and the inverse discrete wavelet transform (IDWT) (8.3.5) with p, q in Sect. 8.3, on p.406, can be written as

$$\mathbf{c} = (\widetilde{\ell}^{\,-} * \mathbf{x}) \downarrow 2, \quad \mathbf{d} = (\widetilde{\mathbf{h}}^{-} * \mathbf{x}) \downarrow 2, \tag{9.5.1}$$
$$\widetilde{\mathbf{x}} = (\mathbf{c} \uparrow 2) * \boldsymbol{\ell} + (\mathbf{d} \uparrow 2) * \mathbf{h}, \tag{9.5.2}$$

where $\downarrow 2$ and $\uparrow 2$ denote the downsampling and upsampling operations, respectively, and for a filter $\mathbf{g} = \{g_k\}$, $\mathbf{g}^{-} = \{g_k^{-}\}$ denotes its time-reverse, given by $g_k^{-} = g_{-k}, k \in \mathbb{Z}$ (see Remark 1 of Sect. 8.4 on p.427). Furthermore, (9.5.1) and (9.5.2) can be written as

$$C(z^2) = \tfrac{1}{2}\widetilde{L}\left(\tfrac{1}{z}\right) X(z) + \tfrac{1}{2}\widetilde{L}(-\tfrac{1}{z}) \, X(-z),$$

$$(9.5.3)$$

$$D(z^2) = \tfrac{1}{2}\widetilde{H}\left(\tfrac{1}{z}\right) X(z) + \tfrac{1}{2}\widetilde{H}(-\tfrac{1}{z}) \, X(-z), \quad z \in \mathbb{C}\backslash\{0\},$$

and

$$\widetilde{X}(z) = L(z) \, C(z^2) + H(z) \, D(z^2), \quad z \in \mathbb{C}\backslash\{0\} \qquad (9.5.4)$$

(see (8.3.8)–(8.3.9) on p. 409). In addition, $\widetilde{p}, \widetilde{q}, p, q$ are biorthogonal if and only if $\widetilde{L}(z), \widetilde{H}(z), L(z), H(z)$ form a perfect reconstruction (PR) filter bank, which, according to Theorem 2 on p.427, is equivalent to

$$\left(M_{L,H}(z)\right)^T M_{\widetilde{L},\widetilde{H}}\left(\frac{1}{z}\right) = 2I_2, \ z \in \mathbb{C}\backslash\{0\}, \qquad (9.5.5)$$

where $M_{L,H}(z)$ and $M_{\widetilde{L},\widetilde{H}}(z)$ are the modulation matrices of $L(z), H(z)$ and $\widetilde{L}(z), \widetilde{H}(z)$, respectively, defined by

$$M_{L,H}(z) = \begin{bmatrix} L(z) & L(-z) \\ H(z) & H(-z) \end{bmatrix}, \ M_{\widetilde{L},\widetilde{H}}(z) = \begin{bmatrix} \widetilde{L}(z) & \widetilde{L}(-z) \\ \widetilde{H}(z) & \widetilde{H}(-z) \end{bmatrix}.$$

For a PR filter bank $L(z), H(z)$ and $\widetilde{L}(z), \widetilde{H}(z)$, one pair of lowpass and highpass filters $\{L, H\}$ or $\{\widetilde{L}, \widetilde{H}\}$ is called the **PR dual** of the other.

For a filter $\mathbf{g} = \{g_k\}$, denote

$$G_e(z) = \sum_k g_{2k} z^k, \quad G_o(z) = \sum_k g_{2k+1} z^k.$$

Then the z-transform $G(z)$ of $\mathbf{g}$ can be written as

$$G(z) = G_e(z^2) + G_o(z^2)z.$$

Definition 1 | **Polyphase matrix** | *The polyphase matrix of a pair of filters* $\{L(z), H(z)\}$ *is defined by*

$$P(z) = \begin{bmatrix} L_e(z) & L_o(z) \\ H_e(z) & H_o(z) \end{bmatrix}.$$

Observe that the modulation matrix $M_{L,H}(z)$ and polyphase matrix of $L(z), H(z)$ have the relation

$$M_{L,H}(z) = P(z^2)\begin{bmatrix} 1 & 1 \\ z & -z \end{bmatrix}.$$

This, together with the fact that

$$
\begin{bmatrix} 1 & z \\ 1 & -z \end{bmatrix} \begin{bmatrix} 1 & 1 \\ \frac{1}{z} & -\frac{1}{z} \end{bmatrix} = 2I_2,
$$

leads to the following theorem.

Theorem 1 | **PR filter bank and polyphase matrices** | *The filter pairs $\{L(z),$ $H(z)\}$ and $\{\widetilde{L}(z), \widetilde{H}(z)\}$ satisfy (9.5.5); that is, they constitute a PR filter bank, if and only if*

$$
P(z)\, \widetilde{P}\left(\frac{1}{z}\right)^T = I_2,   \tag{9.5.6}
$$

where $P(z)$ and $\widetilde{P}(z)$ are the polyphase matrices of $\{L(z), H(z)\}$ and $\{\widetilde{L}(z), \widetilde{H}(z)\}$, respectively.

Recall that a Laurent polynomial $w(z)$ is called a **monomial** if $w(z) = cz^m$ for some $m \in \mathbb{Z}$ and constant $c \neq 0$. From Theorem 1, we have the following corollary (the proof is left as an exercise).

Corollary 1 | **Existence of FIR PR dual and polyphase matrix** | *An FIR filter pair $\{L(z), H(z)\}$ has an FIR PR filter dual if and only if $det(P(z))$ is a monomial.*

Example 1 | **Lazy wavelet transform** | If $P(z) = \widetilde{P}(z) = I_2$, we obtain very simple PR filters, which are called **Lazy filters**, and we denote them by $L^{(0)}(z), H^{(0)}(z)$ and $\widetilde{L}^{(0)}(z), \widetilde{H}^{(0)}(z)$, with

$$
L^{(0)}(z) = \widetilde{L}^{(0)}(z) = 1, \;\; H^{(0)}(z) = \widetilde{H}^{(0)}(z) = z.
$$

The **Lazy wavelet transform**, the decomposition algorithm (or DWT) with $\widetilde{L}^{(0)}(z)$ and $\widetilde{H}^{(0)}(z)$, is to subsample an input signal into the even and odd indexed samples. More precisely, let $\mathbf{x} = \{x_n\}$ be the input signal. Then $\mathbf{c}$ and $\mathbf{d}$ obtained by the wavelet transform (9.5.1) with $\widetilde{L}^{(0)}(z) = 1$ and $\widetilde{H}^{(0)}(z) = z$ (or equivalently, by (8.3.4) on p.406 with $\widetilde{p}(\omega) = \widetilde{L}^{(0)}(e^{-i\omega}) = 1$ and $\widetilde{q}(\omega) = \widetilde{H}^{(0)}(e^{-i\omega}) = e^{-i\omega}$, or $\widetilde{p}_0 = 2, \widetilde{p}_k = 0, k \neq 0$ and $\widetilde{q}_1 = 2, \widetilde{q}_j = 0, j \neq 1$) are given by

$$
\mathbf{c} = \{x_{2n}\}, \;\; \mathbf{d} = \{x_{2n+1}\}.
$$

On the other hand, the **inverse Lazy wavelet transform**, the wavelet reconstruction algorithm (or IDWT) (9.5.2) with $L^{(0)}(z) = 1$ and $H^{(0)}(z) = z$ (or equivalently, (8.3.5) on p.406 with $p(\omega) = \frac{1}{2}L^{(0)}(e^{-i\omega}) = \frac{1}{2}$ and $q(\omega) = \frac{1}{2}H^{(0)}(e^{-i\omega}) = \frac{1}{2}e^{-i\omega}$, or $p_0 = 1, p_k = 0, k \neq 0$ and $q_1 = 1, q_j = 0, j \neq 1$), is to combine $\mathbf{c} = \{c_n\},\ \mathbf{d} = \{d_n\}$ into a signal, with c_n and d_n filling the $2n$-index and $2n+1$-index positions, that is,

$$(\ldots, c_{n-1}, d_{n-1}, c_n, d_n, c_{n+1}, d_{n+1}, \ldots).$$

∎

Definition 2 | **Lifting** | *Let $\{L(z), H(z)\}$ be a pair of FIR filters. Define*

$$L^{\text{new}}(z) = L(z) + G(z^2)H(z), \tag{9.5.7}$$

and

$$H^{\text{new}}(z) = H(z) + S(z^2)L(z), \tag{9.5.8}$$

where $G(z)$, $S(z)$ are Laurent polynomials. Then $L^{\text{new}}(z)$ and $H^{\text{new}}(z)$ are called lifting filters of $L(z)$ and $H(z)$, respectively.

Theorem 2 | **Lifting scheme** | *Let $L(z)$, $H(z)$ be FIR filters and $L^{\text{new}}(z)$ and $H^{\text{new}}(z)$ be their lifting filters defined by (9.5.7) and (9.5.8) for some Laurent polynomials $G(z)$, $S(z)$. Then:*

(i) *$\{L(z), H(z)\}$ has an FIR PR dual $\{\widetilde{L}(z), \widetilde{H}(z)\}$ if and only if $\{L^{\text{new}}(z), H(z)\}$ has an FIR PR dual $\{\widetilde{L}(z), \widetilde{H}^{\text{new}}(z)\}$, with*

$$\widetilde{H}^{\text{new}}(z) = \widetilde{H}(z) - G\left(\frac{1}{z^2}\right)\widetilde{L}(z); \tag{9.5.9}$$

(ii) *$\{L(z), H(z)\}$ has an FIR PR dual $\{\widetilde{L}(z), \widetilde{H}(z)\}$ if and only if $\{L(z), H^{\text{new}}(z)\}$ has an FIR PR dual $\{\widetilde{L}^{\text{new}}(z), \widetilde{H}(z)\}$, with*

$$\widetilde{L}^{\text{new}}(z) = \widetilde{L}(z) - S\left(\frac{1}{z^2}\right)\widetilde{H}(z). \tag{9.5.10}$$

Proof Here we present the proof of (i). The proof of (ii) is similar and it is left as an exercise (see Exercise 2).

To show (i), observe that

$$L^{\text{new}}(z) = L_e(z^2) + L_o(z^2)z + G(z^2)\left(H_e(z^2) + H_o(z^2)z\right)$$

$$= L_e(z^2) + G(z^2)H_e(z^2) + \left(L_o(z^2) + G(z^2)H_o(z^2)\right)z.$$

Thus, the polyphase matrix $P^{\text{new}}(z)$ of $\{L^{\text{new}}(z), H(z)\}$ is given by

$$P^{\text{new}}(z) = \begin{bmatrix} L_e(z) + G(z)H_e(z) & L_o(z) + G(z)H_o(z) \\ H_e(z) & H_o(z) \end{bmatrix} = \begin{bmatrix} 1 & G(z) \\ 0 & 1 \end{bmatrix} P(z),$$

where $P(z)$ is the polyphase matrix of $\{L(z), H(z)\}$. Since $\det\big(P(z)\big) = \det\big(P^{\text{new}}(z)\big)$, we conclude, by Corollary 1, that $\{L(z), H(z)\}$ has an FIR PR dual if and only if $\{L^{\text{new}}(z), H(z)\}$ has an FIR PR dual.

Let $\widetilde{P}(z)$ and $\widetilde{P}^{\text{new}}(z)$ denote the polyphase matrices of $\{\widetilde{L}(z), \widetilde{H}(z)\}$ and $\{\widetilde{L}(z), \widetilde{H}^{\text{new}}(z)\}$, respectively. Then, following similar arguments as above, we have

$$\widetilde{P}^{\text{new}}(z) = \begin{bmatrix} 1 & 0 \\ -G\left(\frac{1}{z}\right) & 1 \end{bmatrix} \widetilde{P}(z).$$

Thus $P(z)\left(\widetilde{P}\left(\frac{1}{z}\right)\right)^T = I_2$ if and only if $P^{\text{new}}(z)\left(\widetilde{P}^{\text{new}}\left(\frac{1}{z}\right)\right)^T = I_2$. Therefore, $\{\widetilde{L}(z), \widetilde{H}^{\text{new}}(z)\}$, with $\widetilde{H}^{\text{new}}(z)$ given by (9.5.9), is a PR dual of $\{L^{\text{new}}(z), H(z)\}$, provided that $\{\widetilde{L}(z), \widetilde{H}(z)\}$ is a PR dual of $\{L(z), H(z)\}$.   ∎

Next, we consider the relation between DWT and IDWT algorithms which have an FIR PR filter bank and those with a lifted FIR PR filter bank. In the following two theorems, $L^{\text{new}}(z)$, $H^{\text{new}}(z)$, $\widetilde{L}^{\text{new}}(z)$ and $\widetilde{H}^{\text{new}}(z)$ refer to the lifting filters defined by (9.5.7), (9.5.8), (9.5.10) and (9.5.9) for some Laurent polynomials $G(z), S(z)$. Let $\mathbf{g} = \{g_k\}$ and $\mathbf{s} = \{s_k\}$ denote the sequences consisting of the coefficients of $G(z)$ and $S(z)$, respectively.

Theorem 3 |**Forward lifting algorithms**| *Suppose that* $\mathbf{c}$ *and* $\mathbf{d}$ *are the lowpass and highpass outputs corresponding to the input sequence* $\mathbf{x}$, *with* $\{\widetilde{L}(z), \widetilde{H}(z)\}$. *Then*:

(i) *the lowpass and highpass outputs* $\mathbf{c}^{\text{new}}$ *and* $\mathbf{d}^{\text{new}}$ *of* $\mathbf{x}$ *with* $\{\widetilde{L}^{\text{new}}(z), \widetilde{H}(z)\}$ *are given by*

$$c_k^{\text{new}} = c_k - \sum_n s_n d_{k-n}, \ d_k^{\text{new}} = d_k, \ k \in \mathbb{Z}; \qquad (9.5.11)$$

(ii) *the lowpass and highpass outputs* $\mathbf{c}^{\text{new}}$ *and* $\mathbf{d}^{\text{new}}$ *of* $\mathbf{x}$ *with* $\{\widetilde{L}(z), \widetilde{H}^{\text{new}}(z)\}$ *are given by*

$$c_k^{\text{new}} = c_k, \ d_k^{\text{new}} = d_k - \sum_n g_n c_{k-n}, \ k \in \mathbb{Z}. \qquad (9.5.12)$$

Proof Here we only give the proof of (i), since the proof of (ii) is similar and is left as an exercise (see Exercise 3).

To prove (i), we start by noticing that the proof is trivial for the case where $\mathbf{d}_k^{\text{new}} = \mathbf{d}_k$, since the corresponding highpass filters $H^{\text{new}}(z)$ and $H(z)$ are the same. For the relation between lowpass outputs, we use the standard notation $C^{\text{new}}(z)$ to denote the z-transform of $\mathbf{c}^{\text{new}}$. Then, with the formulation (9.5.3) for the wavelet decomposition algorithm, we have

$$C^{\text{new}}(z^2) = \frac{1}{2}\widetilde{L}^{\text{new}}\left(\frac{1}{z}\right) X(z) + \frac{1}{2}\widetilde{L}^{\text{new}}(-\frac{1}{z}) X(-z)$$

$$= \frac{1}{2}\left(\widetilde{L}\left(\frac{1}{z}\right) - S(z^2)\widetilde{H}\left(\frac{1}{z}\right)\right)X(z) + \frac{1}{2}\left(\widetilde{L}(-\frac{1}{z}) - S(z^2)\widetilde{H}(-\frac{1}{z})\right)X(-z)$$

$$= \frac{1}{2}\widetilde{L}\left(\frac{1}{z}\right) X(z) + \frac{1}{2}\widetilde{L}(-\frac{1}{z})X(-z) - S(z^2)\frac{1}{2}\left(\widetilde{H}\left(\frac{1}{z}\right) X(z) + \widetilde{H}(-\frac{1}{z})X(-z)\right)$$

$$= C(z^2) - S(z^2)D(z^2).$$

Thus $C^{\text{new}}(z) = C(z) - S(z)D(z)$, or equivalently,

$$\mathbf{c}^{\text{new}} = \mathbf{c} - \mathbf{s} * \mathbf{d}.$$

This yields (9.5.11), as desired. ∎

Theorem 4 | **Backward lifting algorithms** | *For any two given sequences* $\mathbf{c}^{\text{new}}$ *and* $\mathbf{d}^{\text{new}}$,

(i) $\widetilde{\mathbf{x}}$, *obtained by IDWT from* $\mathbf{c}^{\text{new}}$ *and* $\mathbf{d}^{\text{new}}$ *with* $\{L(z), H^{\text{new}}(z)\}$, *is equal to that obtained by IDWT from* $\mathbf{c}$ *and* $\mathbf{d}$ *with* $\{L(z), H(z)\}$, *provided that*

$$d_k = d_k^{\text{new}}, \ c_k = c_k^{\text{new}} + \sum_n s_n d_{k-n}, \ k \in \mathbb{Z}; \qquad (9.5.13)$$

(ii) $\widetilde{\mathbf{x}}$, *obtained by IDWT from* $\mathbf{c}^{\text{new}}$ *and* $\mathbf{d}^{\text{new}}$ *with* $\{L^{\text{new}}(z), H(z)\}$, *is equal to that obtained by IDWT from* $\mathbf{c}$ *and* $\mathbf{d}$ *with* $\{L(z), H(z)\}$, *provided that*

$$c_k = c_k^{\text{new}}, \ d_k = d_k^{\text{new}} + \sum_n g_n c_{k-n}, \ k \in \mathbb{Z}. \qquad (9.5.14)$$

Proof As in the proof of the previous theorem, here we show (i), and the proof of (ii) is left as an exercise. To show (i), we use the formulation (9.5.4) for IDWT (the wavelet reconstruction algorithm). Let $\mathbf{y}$ be the recovered signal (sequence) obtained by IDWT from $\mathbf{c}$ and $\mathbf{d}$ with $\{L(z), H(z)\}$. Then, from (9.5.13), we have

$$\begin{aligned} Y(z) &= L(z)C(z^2) + H(z)D(z^2) \\ &= L(z)\left(C^{\text{new}}(z^2) + S(z^2)D(z^2)\right) + H(z)D(z^2) \\ &= L(z)C^{\text{new}}(z^2) + \left(L(z)S(z^2) + H(z)\right)D(z^2) \\ &= L(z)C^{\text{new}}(z^2) + H^{\text{new}}(z)D^{\text{new}}(z^2) \\ &= \widetilde{X}(z), \end{aligned}$$

as desired, where (9.5.4) is used to obtain the first and the last equalities. ∎

To summarize, the polyphase matrices $\widetilde{P}^{\text{new}}(z)$, $P^{\text{new}}(z)$ of FIR PR filter pairs $\{\widetilde{L}^{\text{new}}(z), \widetilde{H}(z)\}$ and $\{L(z), H^{\text{new}}(z)\}$, lifted from a PR filter bank $\{\widetilde{L}(z), \widetilde{H}(z)\}$, $\{L(z), H(z)\}$, are given by

$$\widetilde{P}^{\text{new}}(z) = \begin{bmatrix} 1 & -S\left(\frac{1}{z}\right) \\ 0 & 1 \end{bmatrix} \widetilde{P}(z), \quad P^{\text{new}}(z) = \begin{bmatrix} 1 & 0 \\ S(z) & 1 \end{bmatrix} P(z).$$

The extra decomposition and reconstruction algorithms (9.5.11) and (9.5.13) are

Forward lifting algorithm: $\mathbf{c}^{\text{new}} = \mathbf{c} - \mathbf{s} * \mathbf{d}$, $\mathbf{d}^{\text{new}} = \mathbf{d}$;   (9.5.15)

Backward lifting algorithm: $\mathbf{d} = \mathbf{d}^{\text{new}}$, $\mathbf{c} = \mathbf{c}^{\text{new}} + \mathbf{s} * \mathbf{d}$.   (9.5.16)

Observe that these two algorithms are very simple. By simply moving one term $\mathbf{s} * \mathbf{d}$ in the equation from one side to the other side, the one can be obtained from the other. Moreover, one can write them down by simply looking at the polyphase matrix $\widetilde{P}^{\text{new}}(z)$.

Similarly, for the lifted FIR PR filter pairs $\{\widetilde{L}(z), \widetilde{H}^{\text{new}}(z)\}$, $\{L^{\text{new}}(z), H(z)\}$, the polyphase matrices are given by

$$\widetilde{P}^{\text{new}}(z) = \begin{bmatrix} 1 & 0 \\ -G\left(\frac{1}{z}\right) & 1 \end{bmatrix} \widetilde{P}(z), \quad P^{\text{new}}(z) = \begin{bmatrix} 1 & G(z) \\ 0 & 1 \end{bmatrix} P(z),$$

and the corresponding forward lifting algorithm (9.5.12) and backward lifting algorithm (9.5.14) are

Forward lifting algorithm: $\mathbf{c}^{\text{new}} = \mathbf{c}$, $\mathbf{d}^{\text{new}} = \mathbf{d} - \mathbf{g} * \mathbf{c}$;   (9.5.17)

Backward lifting algorithm: $\mathbf{c} = \mathbf{c}^{\text{new}}$, $\mathbf{d} = \mathbf{d}^{\text{new}} + \mathbf{g} * \mathbf{c}$.   (9.5.18)

Again, one can write the algorithms down by simply looking at the polyphase matrix $\widetilde{P}^{\text{new}}(z)$.

Multi-step lifting scheme   For a lifted FIR PR filter bank, we can apply more lifting procedures to get further lifted FIR PR filter banks. For $\widetilde{P}(z) = I_2$, the corresponding matrix $\widetilde{P}^{\text{new}}(z)$ will be a product of upper-triangular and lower-triangular matrices of the form

$$\begin{bmatrix} 1 & -S\left(\frac{1}{z}\right) \\ 0 & 1 \end{bmatrix}, \quad \begin{bmatrix} 1 & 0 \\ -G\left(\frac{1}{z}\right) & 1 \end{bmatrix}, \qquad (9.5.19)$$

and the corresponding decomposition algorithms will consist of several repetitions of (9.5.15) and (9.5.17), which can easily be written down from the factorization of $\widetilde{P}^{\text{new}}(z)$. We illustrate this in the following two examples. ∎

Example 2 Suppose that the polyphase matrix $\widetilde{P}^{\text{new}}(z)$ of the analysis filter bank, lifted from the Lazy filter bank, is given by

$$\widetilde{P}^{\text{new}}(z) = \begin{bmatrix} 1 & 0 \\ -G\left(\frac{1}{z}\right) & 1 \end{bmatrix} \begin{bmatrix} 1 & -S\left(\frac{1}{z}\right) \\ 0 & 1 \end{bmatrix},$$

where $G(z)$ and $S(z)$ are Laurent polynomials. Then the forward lifting algorithm for input $\mathbf{x} = (x_k)$ is given as follows:

    Forward lifting algorithm:
        Splitting(Lazy wavelet transform) : $\mathbf{c} = (x_{2k})$, $\mathbf{d} = (x_{2k+1})$;
        Step 1. $\mathbf{c}^{\text{new}} = \mathbf{c} - \mathbf{s} * \mathbf{d}$;
        Step 2. $\mathbf{d}^{\text{new}} = \mathbf{d} - \mathbf{g} * \mathbf{c}^{\text{new}}$.
    $\mathbf{c}^{\text{new}}$ and $\mathbf{d}^{\text{new}}$ are the lowpass and highpass outputs.

The backward lifting algorithm is the reverse of the above forward lifting algorithm:

    Backward lifting algorithm:
        Step 1. $\mathbf{d} = \mathbf{d}^{\text{new}} + \mathbf{g} * \mathbf{c}^{\text{new}}$;
        Step 2. $\mathbf{c} = \mathbf{c}^{\text{new}} + \mathbf{s} * \mathbf{d}$;
        Combining (Inverse Lazy wavelet transform): $\mathbf{x} = (\ldots, c_k, d_k, \ldots)$.
    $\mathbf{x}$ is recovered from its lowpass and highpass outputs $\mathbf{c}^{\text{new}}$ and $\mathbf{d}^{\text{new}}$.

The "Splitting" and "Combining" procedures in the above algorithms are the wavelet decomposition and reconstruction procedures, respectively, with the Lazy filter bank given by $\widetilde{L}(z) = 1$, $\widetilde{H}(z) = z$ and $L(z) = 1$, $H(z) = z$. ∎

Example 3 Suppose that the polyphase matrix $\widetilde{P}^{\text{new}}(z)$ of the analysis filter bank, lifted from the Lazy filter bank, is given by

$$\widetilde{P}^{\text{new}}(z) = \begin{bmatrix} 1 & -S\left(\frac{1}{z}\right) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -G\left(\frac{1}{z}\right) & 1 \end{bmatrix},$$

where $G(z)$ and $S(z)$ are Laurent polynomials. Then the forward lifting algorithm for input $\mathbf{x}$ is given as follows:

Forward lifting algorithm:

Splitting (Lazy wavelet transform): $\mathbf{c} = (x_{2k})$, $\mathbf{d} = (x_{2k+1})$;

Step 1. $\mathbf{d}^{\text{new}} = \mathbf{d} - \mathbf{g} * \mathbf{c}$;

Step 2. $\mathbf{c}^{\text{new}} = \mathbf{c} - \mathbf{s} * \mathbf{d}^{\text{new}}$.

$\mathbf{c}^{\text{new}}$ and $\mathbf{d}^{\text{new}}$ are the lowpass and highpass outputs.

Again, the backward lifting algorithm is the reverse of the above forward lifting algorithm:

Backward lifting algorithm:

Step 1. $\mathbf{c} = \mathbf{c}^{\text{new}} + \mathbf{s} * \mathbf{d}^{\text{new}}$;

Step 2. $\mathbf{d} = \mathbf{d}^{\text{new}} + \mathbf{g} * \mathbf{c}$;

Combining (Inverse Lazy wavelet transform): $\mathbf{x} = (\ldots, c_k, d_k, \ldots)$.

$\mathbf{x}$ is recovered from its lowpass and highpass outputs $\mathbf{c}^{\text{new}}$ and $\mathbf{d}^{\text{new}}$.

∎

So far, we have shown that, from the polyphase matrix $\widetilde{P}(z)$ of a simple FIR PR filter bank such as the Lazy filter bank, we obtain a lifted FIR PR filter bank by (left) multiplying $\widetilde{P}(z)$ with a product of matrices of the form in (9.5.19). Next, we show that the polyphase matrix of an FIR filter pair which has an FIR PR filter dual can be factorized as the product of matrices of the form in (9.5.19).

Theorem 5 | **Factoring wavelet transform into lifting scheme** | *Suppose   the determinant of the polyphase matrix $\widetilde{P}(z)$ of FIR filters $\widetilde{L}(z)$ and $\widetilde{H}(z)$ is a monomial. Then $\widetilde{P}(z)$ can be written as*

$$\widetilde{P}(z) \tag{9.5.20}$$

$$= \begin{bmatrix} c_1 z^{n_1} & 0 \\ 0 & c_2 z^{n_2} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -G_m\left(\frac{1}{z}\right) & 1 \end{bmatrix} \begin{bmatrix} 1 & -S_m\left(\frac{1}{z}\right) \\ 0 & 1 \end{bmatrix} \cdots \begin{bmatrix} 1 & 0 \\ -G_1\left(\frac{1}{z}\right) & 1 \end{bmatrix} \begin{bmatrix} 1 & -S_1\left(\frac{1}{z}\right) \\ 0 & 1 \end{bmatrix},$$

or

$$\widetilde{P}(z) \tag{9.5.21}$$

$$= \begin{bmatrix} 0 & c_1 z^{n_1} \\ -c_2 z^{n_2} & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -G_m\left(\frac{1}{z}\right) & 1 \end{bmatrix} \begin{bmatrix} 1 & -S_m\left(\frac{1}{z}\right) \\ 0 & 1 \end{bmatrix} \cdots \begin{bmatrix} 1 & 0 \\ -G_1\left(\frac{1}{z}\right) & 1 \end{bmatrix} \begin{bmatrix} 1 & -S_1\left(\frac{1}{z}\right) \\ 0 & 1 \end{bmatrix},$$

where c_1, c_2 are nonzero constants, n_1, n_2 are integers, and $G_j(z)$, $S_j(z)$ are Laurent polynomials.

Remark 1 We have the following remarks on the factorizations in (9.5.20) and (9.5.21).

(a) The factorizations in (9.5.20) and (9.5.21) are not unique (see Example 4 below). $G_m(z)$ and/or $S_1(z)$ in (9.5.20) and/or (9.5.21) could be zero, meaning that the second matrix (from the left) in (9.5.20) or (9.5.21) could be an upper-triangular matrix, and the last matrix could be a lower-triangular matrix.

(b) The factorization in (9.5.21) can be transformed into that in (9.5.20), since we can write

$$\begin{bmatrix} 0 & c_1 z^{n_1} \\ -c_2 z^{n_2} & 0 \end{bmatrix} = \begin{bmatrix} c_1 z^{n_1} & 0 \\ 0 & c_2 z^{n_2} \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}.$$

(c) Replacing the original FIR filter pair $\{\widetilde{L}(z), \widetilde{H}(z)\}$ by $\{z^{m_1}\widetilde{L}(z), z^{m_1-2m_2}\widetilde{H}(z)\}$ for some suitable integers m_1 and m_2, we could choose n_1 and n_2 in (9.5.20) and (9.5.21) to be zero.

∎

Proof of Theorem 5   Recall that

$$\widetilde{P}(z) = \begin{bmatrix} \widetilde{L}_e(z) & \widetilde{L}_o(z) \\ \widetilde{H}_e(z) & \widetilde{H}_o(z) \end{bmatrix}.$$

Since the greatest common divisor (gcd) of $\widetilde{L}_e(z)$ and $\widetilde{L}_o(z)$ is also a divisor of $\det(\widetilde{P}(z))$, and $\det(\widetilde{P}(z))$ is a monomial, this gcd must be a monomial.

Assume that length$(\widetilde{L}_o) \geq$ length$(\widetilde{L}_e)$. Then we have

$$\widetilde{L}_o(z) = -S_1\left(\frac{1}{z}\right)\widetilde{L}_e(z) + R_1(z),$$

where $S_1(z)$ and $R_1(z)$ are Laurent polynomials with

$$\text{length}(R_1) < \text{length}(\widetilde{L}_e).$$

$R_1(z)$ is called the remainder of the division. Thus

$$\widetilde{P}(z) \begin{bmatrix} 1 & S_1(\frac{1}{z}) \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} \widetilde{L}_e(z) & R_1(z) \\ \widetilde{H}_e(z) & * \end{bmatrix}.$$

If length$(R_1) \neq 0$ (namely $R_1(z)$ is not a monomial), then $\widetilde{L}_e(z)$ can be written as

$$\widetilde{L}_e(z) = -G_1(\frac{1}{z})R_1(z) + R_2(z),$$

where $G_1(z)$ and $R_2(z)$ are Laurent polynomials with

$$\text{length}(R_2) < \text{length}(R_1).$$

Hence

$$\widetilde{P}(z)\begin{bmatrix} 1 & S_1(\frac{1}{z}) \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ G_1(\frac{1}{z}) & 1 \end{bmatrix} = \begin{bmatrix} R_2(z) & R_1(z) \\ * & * \end{bmatrix}.$$

If length(R_1) = 0, then $R_2 = 0$. Repeating this procedure (if $R_2 \neq 0$), we obtain

$$\widetilde{P}(z)\begin{bmatrix} 1 & S_1(\frac{1}{z}) \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ G_1(\frac{1}{z}) & 1 \end{bmatrix}\cdots\begin{bmatrix} 1 & S_m(\frac{1}{z}) \\ 0 & 1 \end{bmatrix} = \begin{cases} \begin{bmatrix} Y(z) & 0 \\ W_1(z) & W_2(z) \end{bmatrix}, & \text{or} \\ \begin{bmatrix} 0 & Y(z) \\ -W_2(z) & W_1(z) \end{bmatrix}, \end{cases}$$

where $W_1(z)$, $W_2(z)$ are some Laurent polynomials, and $Y(z)$ is the gcd of $\widetilde{L}_e(z)$ and $\widetilde{L}_o(z)$, which is a monomial. Suppose $Y(z) = c_1 z^{n_1}$ for some integer n_1 and constant $c_1 \neq 0$. Taking the determinant of both sides of the above equation, we have

$$\det(\widetilde{P}(z)) = c_1 z^{n_1} W_2(z).$$

Since $\det(\widetilde{P}(z))$ is a monomial, $W_2(z)$ is also a monomial, say $c_2 z^{n_2}$, where $c_2 \neq 0$. Write

$$\begin{bmatrix} c_1 z^{n_1} & 0 \\ W_1(z) & W_2(z) \end{bmatrix} = \begin{bmatrix} c_1 z^{n_1} & 0 \\ W_1(z) & c_2 z^{n_2} \end{bmatrix}$$
$$= \begin{bmatrix} c_1 z^{n_1} & 0 \\ 0 & c_2 z^{n_2} \end{bmatrix}\begin{bmatrix} 1 & 0 \\ \frac{1}{c_2} z^{-n_2} W_1(z) & 1 \end{bmatrix},$$

and write

$$\begin{bmatrix} 0 & c_1 z^{n_1} \\ -W_2(z) & W_1(z) \end{bmatrix} = \begin{bmatrix} 0 & c_1 z^{n_1} \\ -c_2 z^{n_2} & W_1(z) \end{bmatrix}$$
$$= \begin{bmatrix} 0 & c_1 z^{n_1} \\ -c_2 z^{n_2} & 0 \end{bmatrix}\begin{bmatrix} 1 & -\frac{1}{c_2} z^{-n_2} W_1(z) \\ 0 & 1 \end{bmatrix}.$$

Thus $\widetilde{P}(z)$ can be factorized into (9.5.20) or (9.5.21). ∎

Next, we show how to factorize some commonly used orthogonal and biorthogonal wavelet transforms into lifting schemes. The key is to reduce the lengths of $\widetilde{L}_e(z)$ and $\widetilde{L}_o(z)$ by carrying out elementary column operations on the polyphase matrix $\widetilde{P}(z)$. We will use this idea in the following theorem.

Theorem 6 │ **Matrix column operations** │ *Let $A(z)$ and $B(z)$ be two 2×2 matrices. Then*:

(i) *if $B(z)$ is obtained from $A(z)$ by adding $S\left(\frac{1}{z}\right) \times$ column 1 of A to column 2 of A, then*

$$A(z) = B(z) \begin{bmatrix} 1 & -S\left(\frac{1}{z}\right) \\ 0 & 1 \end{bmatrix};$$

(ii) *if $B(z)$ is obtained from $A(z)$ by adding $G\left(\frac{1}{z}\right) \times$ column 2 of A to column 1 of A, then*

$$A(z) = B(z) \begin{bmatrix} 1 & 0 \\ -G\left(\frac{1}{z}\right) & 1 \end{bmatrix}.$$

The proof is left as an exercise (see Exercise 6). ∎

Example 4 $\boxed{\textbf{Haar filter}}$ Let $\widetilde{L}(z) = 1 + z$, $\widetilde{H}(z) = 1 - z$, $L(z) = \frac{1}{2}(1 + z)$, $H(z) = \frac{1}{2}(1 - z)$ be the unnormalized Haar filter bank. Then we can factorize the polyphase matrix $\widetilde{P}(z)$ of $\{\widetilde{L}(z), \widetilde{H}(z)\}$ as follows:

$$\widetilde{P}(z) = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \rightarrow \text{ add}(-1) \times \text{column 1 to column 2}$$

$$\longrightarrow \begin{bmatrix} 1 & 0 \\ 1 & -2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & -2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -\frac{1}{2} & 1 \end{bmatrix}.$$

Thus, by (i) in Theorem 6, we have

$$\widetilde{P}(z) = \begin{bmatrix} 1 & 0 \\ 0 & -2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -\frac{1}{2} & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}.$$

The corresponding forward and backward lifting algorithms for input **x** are given by (see similar lifting algorithms in Sect. 8.3)

Forward lifting algorithm:

Splitting (Lazy wavelet transform): $\mathbf{c} = (x_{2k})$, $\mathbf{d} = (x_{2k+1})$;

Step 1. $\mathbf{c}^{(1)} = \mathbf{c} + \mathbf{d}$;

Step 2. $\mathbf{d}^{(1)} = \mathbf{d} - \frac{1}{2}\mathbf{c}^{(1)}$;

Step 3. $\mathbf{d}^{(2)} = -2\mathbf{d}^{(1)}$.

$\mathbf{c}^{(1)}$ and $\mathbf{d}^{(2)}$ are the lowpass and highpass outputs.

Backward lifting algorithm:

Step 1. $\mathbf{d}^{(1)} = -\dfrac{1}{2}\mathbf{d}^{(2)}$;

Step 2. $\mathbf{d} = \mathbf{d}^{(1)} + \dfrac{1}{2}\mathbf{c}^{(1)}$;

Step 3. $\mathbf{c} = \mathbf{c}^{(1)} - \mathbf{d}$;

Combining (Inverse Lazy wavelet transform): $\mathbf{x} = (\ldots, c_k, d_k, \ldots)$.

$\mathbf{x}$ is recovered from its lowpass and highpass outputs $\mathbf{c}^{(1)}$ and $\mathbf{d}^{(2)}$.

We should note that the factorization of $\widetilde{P}(z)$ is not unique. Indeed, we have

$$
\widetilde{P}(z) = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \;\rightarrow\; \text{add column 2 to column 1}
$$

$$
\longrightarrow \begin{bmatrix} 2 & 1 \\ 0 & -1 \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 1 & \frac{1}{2} \\ 0 & 1 \end{bmatrix}.
$$

Thus, by (ii) in Theorem 6, we have

$$
\widetilde{P}(z) = \begin{bmatrix} 2 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 1 & \frac{1}{2} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix}.
$$

∎

Example 5 $\boxed{D_4\text{-filter}}$ Let $p(\omega)$ and $q(\omega)$ be the D_4 orthogonal filters given by (8.2.30) and (8.2.31) on p.399. Consider $\widetilde{L}(z)$, $\widetilde{H}(z)$, given by $\widetilde{L}(e^{-i\omega}) = p(\omega)$, $\widetilde{H}(e^{-i\omega}) = e^{i2\omega}q(\omega)$. Observe that we have shifted the coefficients of $q(\omega)$ two units to the left to define $\widetilde{H}(z)$. Thus

$$
\widetilde{L}(z) = \frac{1-\sqrt{3}}{8} + \frac{3-\sqrt{3}}{8}z + \frac{3+\sqrt{3}}{8}z^2 + \frac{1+\sqrt{3}}{8}z^3,
$$

$$
\widetilde{H}(z) = \frac{1+\sqrt{3}}{8}z^{-2} - \frac{3+\sqrt{3}}{8}z^{-1} + \frac{3-\sqrt{3}}{8} - \frac{1-\sqrt{3}}{8}z.
$$

The polyphase matrix $\widetilde{P}(z)$ of $\{\widetilde{L}(z), \widetilde{H}(z)\}$ is factorized as follows:

$$
8\widetilde{P}(z) = \begin{bmatrix} 1-\sqrt{3}+(3+\sqrt{3})z & 3-\sqrt{3}+(1+\sqrt{3})z \\ (1+\sqrt{3})z^{-1}+3-\sqrt{3} & -(3+\sqrt{3})z^{-1}-1+\sqrt{3} \end{bmatrix}
$$

add $(-\sqrt{3}) \times$ column 2 to column 1

$$
\longrightarrow \begin{bmatrix} 4-4\sqrt{3} & 3-\sqrt{3}+(1+\sqrt{3})z \\ (4+4\sqrt{3})z^{-1} & -(3+\sqrt{3})z^{-1}-1+\sqrt{3} \end{bmatrix}
$$

$$\text{add } \left(\frac{\sqrt{3}}{4} + \frac{2+\sqrt{3}}{4}z\right) \times \text{column 1 to column 2}$$

$$\longrightarrow \begin{bmatrix} 4 - 4\sqrt{3} & 0 \\ (4 + 4\sqrt{3})z^{-1} & 4 + 4\sqrt{3} \end{bmatrix}$$

$$= \begin{bmatrix} 4 - 4\sqrt{3} & 0 \\ 0 & 4 + 4\sqrt{3} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ z^{-1} & 1 \end{bmatrix}.$$

Thus, by Theorem 6, we have

$$\widetilde{P}(z) = \begin{bmatrix} \frac{1-\sqrt{3}}{2} & 0 \\ 0 & \frac{1+\sqrt{3}}{2} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ z^{-1} & 1 \end{bmatrix} \begin{bmatrix} 1 & -\frac{\sqrt{3}}{4} - \frac{2+\sqrt{3}}{4}z \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ \sqrt{3} & 1 \end{bmatrix}.$$

The corresponding forward and backward lifting algorithms for input $\mathbf{x}$ are given as follows:

Forward lifting algorithm:

Splitting (Lazy wavelet transform): $\mathbf{c} = (x_{2k})$, $\mathbf{d} = (x_{2k+1})$;

Step 1. $d_k^{(1)} = d_k + \sqrt{3}c_k$;

Step 2. $c_k^{(1)} = c_k - \frac{\sqrt{3}}{4}d_k^{(1)} - \frac{2+\sqrt{3}}{4}d_{k+1}^{(1)}$;

Step 3. $d_k^{(2)} = d_k^{(1)} + c_{k-1}^{(1)}$;

Step 4. $c_k^{(2)} = \frac{1-\sqrt{3}}{2}c_k^{(1)}, d_k^{(3)} = \frac{1+\sqrt{3}}{2}d_k^{(2)}$.

$\mathbf{c}^{(2)}$ and $\mathbf{d}^{(3)}$ are the lowpass and highpass outputs.

Backward lifting algorithm:

Step 1. $c_k^{(1)} = -(1 + \sqrt{3})c_k^{(2)}, d_k^{(2)} = (\sqrt{3} - 1)d_k^{(3)}$;

Step 2. $d_k^{(1)} = d_k^{(2)} - c_{k-1}^{(1)}$;

Step 3. $c_k = c_k^{(1)} + \frac{\sqrt{3}}{4}d_k^{(1)} + \frac{2+\sqrt{3}}{4}d_{k+1}^{(1)}$;

Step 4. $d_k = d_k^{(1)} - \sqrt{3}c_k$;

Combining (Inverse Lazy wavelet transform): $\mathbf{x} = (\ldots, c_k, d_k, \ldots)$.

$\mathbf{x}$ is recovered from its lowpass and highpass outputs $\mathbf{c}^{(2)}$ and $\mathbf{d}^{(3)}$. ■

Example 6 $\boxed{5/3\text{-tap biorthogonal filters}}$ Let $\widetilde{p} = \{\widetilde{p}_k\}, \widetilde{q} = \{\widetilde{q}_k\}$ be the FIR filter pair of the 5/3-biorthogonal filter bank given in Example 5 on p.402. With $z = e^{-i\omega}$, denote $\widetilde{L}(z) = \widetilde{p}(\omega)$, $\widetilde{H}(z) = \widetilde{q}(\omega)$, that is,

$$\widetilde{L}(z) = -\frac{1}{8}\frac{1}{z} + \frac{1}{4} + \frac{3}{4}z + \frac{1}{4}z^2 - \frac{1}{8}z^3,$$

$$\widetilde{H}(z) = -\frac{1}{4}\frac{1}{z} + \frac{1}{2} - \frac{1}{4}z.$$

The polyphase matrix $\widetilde{P}(z)$ of $\{\widetilde{L}(z), \widetilde{H}(z)\}$ is factorized as follows:

$$2\widetilde{P}(z) = \begin{bmatrix} \frac{1}{2} + \frac{1}{2}z & \frac{3}{2} - \frac{1}{4}\left(\frac{1}{z} + z\right) \\ 1 & -\frac{1}{2}\left(\frac{1}{z} + 1\right) \end{bmatrix}$$

$$\text{add}\left(\frac{1}{2} + \frac{1}{2z}\right) \times \text{column 1 to column 2}$$

$$\longrightarrow \begin{bmatrix} \frac{1}{2} + \frac{1}{2}z & 2 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 2 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ \frac{1}{4}(1+z) & 1 \end{bmatrix}.$$

Thus, by (i) in Theorem 6, we have

$$\widetilde{P}(z) = \begin{bmatrix} 0 & 1 \\ 2 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ \frac{1}{4}(1+z) & 1 \end{bmatrix} \begin{bmatrix} 1 & -\frac{1}{2}(1 + \frac{1}{z}) \\ 0 & 1 \end{bmatrix}.$$

The corresponding forward and backward lifting algorithms for input **x** are:

Forward lifting algorithm:

Splitting (Lazy wavelet transform): $\mathbf{c} = (x_{2k})$, $\mathbf{d} = (x_{2k+1})$;

Step 1. $c_k^{(1)} = c_k - \frac{1}{2}(d_k + d_{k-1})$;

Step 2. $d_k^{(1)} = d_k + \frac{1}{4}(c_k^{(1)} + c_{k+1}^{(1)})$;

Step 3. $c_k^{(2)} = d_k^{(1)}, d_k^{(2)} = \frac{1}{2}c_k^{(1)}$.

$\mathbf{c}^{(2)}$ and $\mathbf{d}^{(2)}$ are the lowpass and highpass outputs.

Backward lifting algorithm:

Step 1. $c_k^{(1)} = 2d_k^{(2)}, d_k^{(1)} = c_k^{(2)}$;

Step 2. $d_k = d_k^{(1)} - \frac{1}{4}(c_k^{(1)} + c_{k+1}^{(1)})$;

Step 3. $c_k = c_k^{(1)} + \dfrac{1}{2}(d_k + d_{k-1})$;

Combining (Inverse Lazy wavelet transform): $\mathbf{x} = (\ldots, c_k, d_k, \ldots)$.

$\mathbf{x}$ is recovered from its lowpass and highpass outputs; $\mathbf{c}^{(2)}$ and $\mathbf{d}^{(2)}$.

■

Example 7 $\boxed{\text{9/7-tap biorthogonal filters}}$ Let $\widetilde{p} = \{\widetilde{p}_k\}, \widetilde{q} = \{\widetilde{q}_k\}$ be the FIR filter pair of the 9/7-biorthogonal filter bank constructed in Sect. 9.4 of Chap. 9, on p.477. With $z = e^{-i\omega}$, denote $\widetilde{L}(z) = e^{i2\omega}\widetilde{p}(\omega)$, $\widetilde{H}(z) = -e^{i2\omega}\widetilde{q}(\omega)$, that is,

$$
\begin{aligned}
\widetilde{L}(z) &= \frac{1}{z^2 2^4}(1+z)^4 (-\frac{5\theta}{4})\left(z^2 + \frac{1}{z^2} - (6+4\theta)(z + \frac{1}{z}) + 10 + 8\theta - \frac{4}{5\theta}\right) \\
&= \widetilde{\ell}_0 + \widetilde{\ell}_1(z + \frac{1}{z}) + \widetilde{\ell}_2(z^2 + \frac{1}{z^2}) + \widetilde{\ell}_3(z^3 + \frac{1}{z^3}) + \widetilde{\ell}_4(z^4 + \frac{1}{z^4}), \\
\widetilde{H}(z) &= \frac{1}{2^4}z^3(1 - \frac{1}{z})^4 \frac{1}{4\theta}\left(-z - \frac{1}{z} - (2 - 4\theta)\right) \\
&= \widetilde{h}_1 z + \widetilde{h}_2(z^2 + 1) + \widetilde{h}_3(z^3 + \frac{1}{z}) + \widetilde{h}_4(z^4 + \frac{1}{z^2}),
\end{aligned}
$$

where θ ($\approx -0.3423840948583691$) is an irrational number given by (9.3.23) on p.462, and

$$
\begin{aligned}
&\widetilde{\ell}_0 = \frac{3}{8} - \frac{35}{32}\theta - \frac{5}{4}\theta^2, \ \widetilde{\ell}_1 = \frac{1}{4} - \frac{5}{32}\theta - \frac{5}{16}\theta^2, \\
&\widetilde{\ell}_2 = \frac{1}{16} + \frac{5}{8}\theta + \frac{5}{8}\theta^2, \ \widetilde{\ell}_3 = \frac{5}{32}\theta + \frac{5}{16}\theta^2, \ \widetilde{\ell}_4 = -\frac{5}{64}\theta, \\
&\widetilde{h}_1 = \frac{3}{8} - \frac{1}{16\theta}, \ \widetilde{h}_2 = \frac{1}{64\theta} - \frac{1}{4}, \ \widetilde{h}_3 = \frac{1}{16} + \frac{1}{32\theta}, \ \widetilde{h}_4 = -\frac{1}{64\theta}.
\end{aligned}
$$

The numerical values of $\widetilde{\ell}_j, \widetilde{h}_j$ can be obtained by

$$
\widetilde{\ell}_j = \frac{1}{2}\widetilde{p}_{j+2}, \ \widetilde{h}_j = -\frac{1}{2}\widetilde{q}_{j+2},
$$

with $\widetilde{p}_j, \widetilde{q}_j$ given on p.478.

To factorize the polyphase matrix $\widetilde{P}(z)$ of $\{\widetilde{L}(z), \widetilde{H}(z)\}$, we reduce the lengths of the (1, 1)-entry and the (1, 2)-entry of $\widetilde{P}(z)$. In the following, we use $A[j, k]$ to denote the (j, k)-entry of the 2×2 matrix $A(z)$ of Laurent polynomials, and we use $\text{coeff}(A[j, k], z^n)$ to denote its coefficient with power z^n. Then, we have

$$
\widetilde{P}(z) = \begin{bmatrix} \widetilde{\ell}_0 + \widetilde{\ell}_2(z + \frac{1}{z}) + \widetilde{\ell}_4(z^2 + \frac{1}{z^2}) & \widetilde{\ell}_1(1 + \frac{1}{z}) + \widetilde{\ell}_3(z + \frac{1}{z^2}) \\ \widetilde{h}_2(1 + z) + \widetilde{h}_4(z^2 + \frac{1}{z}) & \widetilde{h}_1 + \widetilde{h}_3(z + \frac{1}{z}) \end{bmatrix}.
$$

With

$$a = -\mathrm{coeff}(\widetilde{P}[1,1], z^2) \Big/ \mathrm{coeff}(\widetilde{P}[1,2], z) = -\widetilde{\ell}_4 \Big/ \widetilde{\ell}_3,$$

we add $a(1+z)\times$ column 2 to column 1 of $\widetilde{P}(z)$ to obtain

$\widetilde{P}(z)$

$$\longrightarrow A(z) = \begin{bmatrix} *+*(z+\frac{1}{z}) & \widetilde{\ell}_1(1+\frac{1}{z})+\widetilde{\ell}_3(z+\frac{1}{z^2}) \\ *(1+z) & \widetilde{h}_1+\widetilde{h}_3(z+\frac{1}{z}) \end{bmatrix}$$

with $b = -\dfrac{\mathrm{coeff}(A[1,2], z)}{\mathrm{coeff}(A[1,1], z)}, addb(1+\dfrac{1}{z}) \times$ column 1 to column 2 of A(z)

$$\longrightarrow B(z) = \begin{bmatrix} *+*(z+\frac{1}{z}) & *(1+\frac{1}{z}) \\ *(1+z) & * \end{bmatrix}$$

with $c = -\dfrac{\mathrm{coeff}(B[1,1], z)}{\mathrm{coeff}(B[1,2], z^0)}$, add $c(1+z) \times$ column 2 to column 1   of $B(z)$

$$\longrightarrow C(z) = \begin{bmatrix} g & *(1+\frac{1}{z}) \\ 0 & \frac{1}{2g} \end{bmatrix} = \begin{bmatrix} g & 0 \\ 0 & \frac{1}{2g} \end{bmatrix} \begin{bmatrix} 1 & -d(1+\frac{1}{z}) \\ 0 & 1 \end{bmatrix},$$

where

$$g = \frac{1}{4} - \frac{5}{2}\theta^2 - \frac{5}{2}\theta \approx 0.8128930661159611,$$

$$d = -\frac{3+2\theta}{11+34\theta+50\theta^2} \approx -0.4435068520439712.$$

The constants a, b, c, d, g can be calculated by using computer software, such as Maple. The expressions and numerical values of a, b, c are given by

$$a = \frac{1}{2(1+2\theta)} \approx 1.5861343420599236,$$

$$b = \frac{1-\theta-10\theta^2}{4(1+4\theta+10\theta^2)} \approx 0.0529801185729614,$$

$$c = \frac{1+6\theta-2\theta^2}{2(1+2\theta)(3+2\theta)} \approx -0.8829110755309333$$

(we have used the fact that $1+4\theta+10\theta^2+20\theta^3 = 0$ to simplify b, c, d). Thus, $\widetilde{P}(z)$ can be factorized as

$$\widetilde{P}(z) = \begin{bmatrix} g & 0 \\ 0 & \frac{1}{2g} \end{bmatrix} \begin{bmatrix} 1 & -d(1+\frac{1}{z}) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -c(1+z) & 1 \end{bmatrix} \begin{bmatrix} 1 & -b(1+\frac{1}{z}) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -a(1+z) & 1 \end{bmatrix}.$$

The corresponding forward and backward lifting algorithms for input $\mathbf{x}$ are given as follows:

Forward lifting algorithm:

Splitting (Lazy wavelet transform): $\mathbf{c} = (x_{2k}), \mathbf{d} = (x_{2k+1})$;

> **Step 1.** $d_k^{(1)} = d_k - a(c_k + c_{k+1})$;
>
> **Step 2.** $c_k^{(1)} = c_k - b(d_k^{(1)} + d_{k-1}^{(1)})$;
>
> **Step 3.** $d_k^{(2)} = d_k^{(1)} - c(c_k^{(1)} + c_{k+1}^{(1)})$;
>
> **Step 4.** $c_k^{(2)} = c_k^{(1)} - d(d_k^{(2)} + d_{k-1}^{(2)})$;
>
> **Step 5.** $c_k^{(3)} = g\, c_k^{(2)}, d_k^{(3)} = \dfrac{1}{2g} d_k^{(2)}$.

$\mathbf{c}^{(3)}$ *and* $\mathbf{d}^{(3)}$ are the lowpass and highpass outputs.

Backward lifting algorithm:

> **Step 1.** $c_k^{(2)} = \dfrac{1}{g} c_k^{(3)}, d_k^{(2)} = 2g\, d_k^{(3)}$;
>
> **Step 2.** $c_k^{(1)} = c_k^{(2)} + d(d_k^{(2)} + d_{k-1}^{(2)})$;
>
> **Step 3.** $d_k^{(1)} = d_k^{(2)} + c(c_k^{(1)} + c_{k+1}^{(1)})$;
>
> **Step 4.** $c_k = c_k^{(1)} + b(d_k^{(1)} + d_{k-1}^{(1)})$;
>
> **Step 5.** $d_k = d_k^{(1)} + a(c_k + c_{k+1})$;
>
> **Combining** (Inverse Lazy wavelet transform): $\mathbf{x} = (\ldots, c_k, d_k, \ldots)$.

$\mathbf{x}$ is recovered from its lowpass and highpass outputs $\mathbf{c}^{(3)}$ and $\mathbf{d}^{(3)}$.

■

Exercises

Exercise 1 Give a rigorous proof of Corollary 1.
   Hint: You may apply Theorem 1 in this section and Theorem 3 on p.427.

Exercise 2 Let $H^{\text{new}}(z)$ and $\widetilde{L}^{\text{new}}(z)$ be the lifting filters of $H(z)$ and $\widetilde{L}(z)$, defined by (9.5.8) and (9.5.10), respectively. Show that

(a) $\{L(z), H(z)\}$ has an FIR PR dual if and only if $\{L(z), H^{\text{new}}(z)\}$ has an FIR PR dual;

(b) $\{\widetilde{L}^{\text{new}}(z), \widetilde{H}(z)\}$ is a PR dual to $\{L(z), H^{\text{new}}(z)\}$ if $\{\widetilde{L}(z), \widetilde{H}(z)\}$ is a PR dual of $\{L(z), H(z)\}$.

Exercise 3 Prove (ii) in Theorem 3.

Exercise 4 Prove (ii) in Theorem 4.

Exercise 5 Suppose that the FIR filter pair $\{\widetilde{L}(z), \widetilde{H}(z)\}$ has an FIR PR filter dual. Show that $\{z^{m_1} \widetilde{L}(z), z^{m_1-2m_2} \widetilde{H}(z)\}$ also has an FIR PR filter dual, for some integers m_1 and m_2 different from 0.

Exercise 6 Prove Theorem 6.

Exercise 7 Let

$$L(z) = -\frac{1}{8}\frac{1}{z^2} + \frac{1}{8}\frac{1}{z} + 1 + z + \frac{1}{8}z^2 - \frac{1}{8}z^3;$$
$$H(z) = -1 + z.$$

(a) Find the polyphase matrix $P(z)$ of $\{L(z), H(z)\}$.
(b) Compute the determinant of $P(z)$.
(c) Decide whether $\{L(z), H(z)\}$ has an FIR PR dual or not.

Exercise 8 Repeat Exercise 7 for

$$L(z) = -\frac{1}{8}\frac{1}{z^2} + \frac{1}{8}\frac{1}{z} + 1 + z + \frac{1}{8}z^2 - \frac{1}{8}z^3,$$
$$H(z) = -\frac{1}{z} + 1.$$

Exercise 9 Let

$$L(z) = z + \frac{1}{2}(1 + z^2);$$
$$H(z) = \frac{3}{2} - \frac{1}{2}(z + \frac{1}{z}) - \frac{1}{4}(z^2 + \frac{1}{z^2}).$$

(a) Find the polyphase matrix $P(z)$ of $\{L(z), H(z)\}$.
(b) Show that $\{L(z), H(z)\}$ has an FIR PR dual.

Exercise 10 Repeat Exercise 9 for

$$L(z) = \frac{5}{2} + \frac{5}{16}(z + \frac{1}{z}) - \frac{3}{4}(z^2 + \frac{1}{z^2}) + \frac{3}{16}(z^3 + \frac{1}{z^3}),$$
$$H(z) = \frac{3}{4}z - \frac{1}{2}(1 + z^2) + \frac{1}{8}\left(\frac{1}{z} + z^3\right).$$

Exercise 11 Factorize the polyphase matrix $P(z)$ of $\{L(z), H(z)\}$ as given in Exercise 9, and then provide the corresponding forward and backward lifting algorithms.

Exercise 12 Repeat Exercise 11 by using both $L(z)$ and $H(z)$ as given in Exercise 10.

Chapter 10
Wavelet Analysis
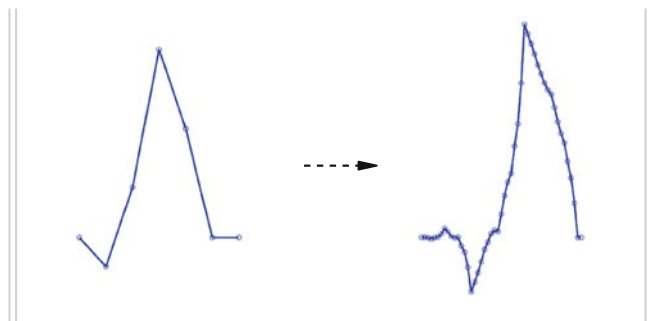
The transition operator T_p, along with its representation matrix in terms of a given finite sequence p, as introduced and studied in Sect. 9.1 of Chap. 9, is again applied to our study of scaling functions and wavelets in this chapter. In particular, while the characterization of the properties of stability and orthogonality of the refinable function ϕ was formulated in terms of the Gramian function $G_\phi(\omega)$ in Chap. 9, these properties will be characterized in terms of the properties of the eigenvalues and 1-eigenvectors of the transition operator T_p in this chapter. This consideration is necessary, since there is no explicit expression of the refinable function ϕ in the formulation of $G_\phi(\omega)$ in general.

When the sequence $p = \{p_k\}_{k=0}^N$ satisfies the (necessary) condition that $\sum_k p_k = 2$, we will first prove in Sect. 10.1 that the infinite product $\prod_{j=1}^\infty p(\frac{\omega}{2^j})$ converges absolutely for each ω. Here and throughout Chap. 10, we will use the notation p, without boldface, for any finite refinement sequence (see Remark 1 in Sect. 8.2 on p. 396), and $p(\omega)$ to denote the two-scale symbol of p.

By the Paley-Wiener Theorem, it can be shown that the above infinite product converges to an entire function, which is the Fourier transform of some compactly supported distribution ϕ. One of the main objectives of this first section is to show that a necessary and sufficient condition for the distribution ϕ to be a function in $L_2(\mathbb{R})$ is that the transition operator T_p has a 1-eigenfunction $v(\omega)$, which is a non-negative trigonometric polynomial of degree given by the length N of the support of the refinement sequence p, such that $v(0) > 0$. Other topics of investigation in Sect. 10.1 include the support of ϕ, the sum-rule and the property of partition of unity.

Characterizations of stability and orthogonality of refinable functions ϕ in terms of the transition operators, defined by the associated refinement sequences p, is the subject of investigation in Sect. 10.2. The Condition E of T_p and the sum-rule order of p, defined in Sect. 9.3 of Chap. 9, are used for the characterization. For example, it will be shown that a refinable function ϕ is stable, if and only if (i) the corresponding

refinement sequence p has sum-rule order ≥ 1, (ii) the transition operator T_p satisfies Condition E, and (iii) there exists a 1-eigenfunction $v_0(\omega)$ of T_p that satisfies either $v_0(\omega) > 0$, or $v_0(\omega) < 0$, for all $\omega \in \mathbb{R}$. Another result is that a compactly supported refinable function ϕ is orthogonal, if and only if (i) the corresponding refinement sequence p is a QMF, (ii) it has sum-rule order ≥ 1, and (iii) its transition operator T_p satisfies Condition E.

With only very few exceptions such as the Cardinal B-splines, compactly supported scaling functions (i.e. stable refinable functions) do not have explicit expressions. For instance, again with the exception of B-splines (such as the Haar scaling function), the compactly supported orthogonal and biorthogonal scaling functions and wavelets constructed in Chap. 9, with demonstrative examples in Chap. 8, are only formulated in terms of the refinement and two-scale equations. Of course, the values of a refinable function $\phi(x)$ at the dyadic points $x = k/2^j$, for $k = 0, \ldots, 2^j N$ and $j = 0, 1, \ldots, n$, can be computed iteratively for any desirable positive integer n, by first finding the 1-eigenvector $\mathbf{y}_0 = [\phi(0), \phi(1), \ldots, \phi(N)]^T$ of the $(N + 1)$-dimensional square matrix $[p_{2j-k}]$, $0 \leq j, k \leq N$, and then by applying the refinement equation to compute $\phi(k/2^j)$ for $j = 1, \ldots, n$, and finally by applying the two-scale relation of the wavelet to compute $\psi(k/2^n)$ from $\phi(k/2^{n-1})$. The objective of Sect. 10.3 is to introduce and study the convergence of the cascade algorithm, for computing a sequence of functions $\phi_n(x)$, in terms of some linear combinations of any desirable compactly supported initial function $\phi_0(x)$, that approximates $\phi(x)$ for all $x \in \mathbb{R}$ and sufficiently large n. The main result of this section is that convergence of the cascade algorithm is guaranteed if and only if the refinement sequence p has sum-rule order ≥ 1 and its transition operator satisfies Condition E. We will also apply this convergence result to prove the characterization for the biorthogonality of scaling functions.

As mentioned above, the compactly supported scaling functions ϕ and wavelets ψ, that are orthogonal or biorthogonal, do not have explicit expressions in general. On the other hand, to determine the order of smoothness of those functions that are written as (finite) linear combinations of $\phi(2^m x - k)$ and $\psi(2^j x - k)$, for $j, k, m \in \mathbb{Z}$, it is necessary (and sufficient) to know the order of smoothness of the scaling function ϕ (since the wavelet $\psi(x)$ is a linear combination of $\phi(2x - k)$ from the wavelet two-scale relation). Section 10.4 of this chapter is devoted to the study of the Sobolev smoothness of ϕ. In this book, the order of Sobolev smoothness will be given in terms of the eigenvalues of the transition operator T_p associated with the refinement sequence p of ϕ. To introduce this concept, we need the notion of the Sobolev space.

For any $\gamma \geq 0$, a function f is said to be in the Sobolev space $\mathbb{W}^\gamma(\mathbb{R})$, if f satisfies the condition $(1 + \omega^2)^{\frac{\gamma}{2}} \widehat{f}(\omega) \in L_2(\mathbb{R})$. Then the order of Sobolev smoothness is defined by the critical Sobolev exponent

$$\nu_f = \sup\{\gamma : f \in \mathbb{W}^\gamma(\mathbb{R})\}.$$

To describe the results in this section, the concept of spectral radius for square matrices, studied in Chap. 3, is extended to the transition operator T_p and its restric-

tion on the subspaces $\mathbb{U}_{2L}$ of trigonometric polynomials $u(\omega)$ of degree N whose derivatives up to order $2L - 1$ vanish at $\omega = 0$. Here, N is the length of the support of the refinement sequence p and L denotes the sum-rule order of p. The natural extension of the notion of "spectral radius" from a square matrix A, studied in Chap. 3, to a more general operator T, is to apply the result $\rho(A) = \lim_{n \to \infty} |||A^n|||^{\frac{1}{n}}$ in Theorem 3 of Sect. 3.2 to define the spectral radius of T, namely: $\rho(T) = \lim_{n \to \infty} |||T^n|||^{\frac{1}{n}}$. In Sect. 10.4 of this chapter, we will consider the operator $T = T_p|_{\mathbb{U}_{2L}}$ and denote $\rho_0 = \rho(T_p|_{\mathbb{U}_{2L}})$. In addition, let $\sigma(\mathcal{T}_p)$ denote the set of the eigenvalues of the representation matrix $\mathcal{T}_p$ of the transition operator T_p. The main results in this section are: firstly, the estimate of the critical Sobolev exponent

$$\nu_\phi \geq -\frac{\log \rho_0}{2 \log 2},$$

where this inequality can be replaced by equality, if the refinable function ϕ is stable (i.e., ϕ is a scaling function); and secondly,

$$\rho_0 = max\left\{|\lambda| : \lambda \in \sigma(\mathcal{T}_p) \backslash \{1, \frac{1}{2}, \dots, \frac{1}{2^{2L-1}}\}\right\}.$$

10.1 Existence of Refinable Functions in $L_2(\mathbb{R})$

The first objective of this section is to establish an absolute convergence result, to be stated in Theorem 1, for the infinite product $\prod_{j=1}^{\infty} p(\frac{\omega}{2^j})$, associated with an arbitrary finite sequence $p = \{p_k\}$ with $\sum_k p_k = 2$. Here and throughout, $p(\omega)$ denotes the two-scale symbol of p, as defined in (9.1.1) in Chap. 9, namely:

$$p(\omega) = \frac{1}{2} \sum_k p_k e^{-ik\omega}.$$

Since $p(0) = \frac{1}{2} \sum_k p_k = 1$, the above infinite product converges to 1 at $\omega = 0$, so that the limit function, denoted by $\widehat{\phi}(\omega)$, satisfies the normalization condition $\widehat{\phi}(0) = 1$. The second objective of this section is to apply the Fundamental Lemma of transition operators (i.e. Theorem 3) in Sect. 9.1 of Chap. 9 to show that the limit function $\widehat{\phi}(\omega)$ is the Fourier transform of some function $\phi \in L_2(\mathbb{R})$, under certain assumptions on the transition operator T_p associated with the given finite refinement sequence (mask) p. This result will be established in Theorem 3 in this section. The third objective of this section is to derive various properties of the function ϕ, including: the support of ϕ as dictated by the support of its refinement mask p, the property of partition of unity (PU), and the sum-rule property. Let us first recall (see Sect. 8.2) that for a finite sequence $p = \{p_k\}_{k=0}^N$, the solution ϕ (if it exists) of the refinement equation

$$\phi(x) = \sum_{k=0}^{N} p_k \phi(2x - k) \tag{10.1.1}$$

is called the normalized solution, if the principal value of its integral over $\mathbb{R}$ is equal to 1 (i.e. $\widehat{\phi}(0) = 1$). An equivalent formulation of (10.1.1) is

$$\widehat{\phi}(\omega) = p(\frac{\omega}{2})\widehat{\phi}(\frac{\omega}{2}), \tag{10.1.2}$$

and

$$\widehat{\phi}(\omega) = \prod_{j=1}^{\infty} p(\frac{\omega}{2^j}), \tag{10.1.3}$$

as already mentioned in (8.2.21) on p.397 in Sect. 8.2 of Chap. 8. Of course, the above statements are meaningful, only if the infinite product in (10.1.3) converges to the Fourier transform of the normalized solution ϕ (again, if it exists) of the refinement equation (10.1.1). First let us show that the infinite product (10.1.3) always converges, as long as the finite sequence p sums to 2, as follows.

Theorem 1 $\boxed{\text{Pointwise convergence of } \prod_{j=1}^{\infty} p(\frac{\omega}{2^j})}$ *Let* $p = \{p_k\}$ *be any finite sequence with* $\sum_k p_k = 2$. *Then* $\prod_{j=1}^{\infty} p(\frac{\omega}{2^j})$ *converges absolutely for each* $\omega \in \mathbb{R}$ *in the sense that* $B_n = |\prod_{j=1}^{n} p(\frac{\omega}{2^j})|$ *converges for each* $\omega \in \mathbb{R}$.

Proof Since $\ln B_n = \sum_{j=1}^{n} \ln |p(\frac{\omega}{2^j})|$, it is sufficient to show that $\sum_{j=1}^{\infty} \ln |p(\frac{\omega}{2^j})|$ converges pointwise on $\mathbb{R}$.

To prove the convergence of this infinite series, we observe that since $p(\omega)$ is a trigonometric polynomial, $p'(\omega)$ is also a trigonometric polynomial; and being periodic and continuous, it is bounded by some positive constant C, namely: $|p'(\omega)| \leq C$ for all $\omega \in \mathbb{R}$. Therefore, by the Mean Value Theorem, we have

$$|p(\omega) - 1| = |p(\omega) - p(0)| = |p'(\xi)(\omega - 0)| \leq C|\omega|,$$

for some ξ. Hence, we have $|p(\frac{\omega}{2^j}) - 1| \leq \frac{C|\omega|}{2^j}$, so that

$$1 - \frac{C|\omega|}{2^j} \leq |p(\frac{\omega}{2^j})| \leq 1 + \frac{C|\omega|}{2^j}, \tag{10.1.4}$$

for all integers $j \geq 1$.

On the other hand, in view of the inequalities $\ln(1 + x) \leq x$, for $x \geq 0$; and

$$-2x \leq \ln(1 - x), \text{ for } 0 \leq x \leq \frac{1}{2} \tag{10.1.5}$$

(see Exercise 1), we may conclude that for an arbitrary given $\omega \in \mathbb{R}$, there exists some positive integer J, such that

$$\frac{C|\omega|}{2^j} \leq \frac{1}{2}, \text{ for } j \geq J.$$

Thus, for $j \geq J$, by (10.1.4) and (10.1.5),

$$-2 \cdot \frac{C|\omega|}{2^j} \leq \ln(1 - \frac{C|\omega|}{2^j}) \leq \ln|p(\frac{\omega}{2^j})| \leq \ln(1 + \frac{C|\omega|}{2^j}) \leq \frac{C|\omega|}{2^j}.$$

Therefore,

$$\sum_{j=1}^{\infty} \left|\ln|p(\frac{\omega}{2^j})|\right| \leq \sum_{j=1}^{J-1} \left|\ln|p(\frac{\omega}{2^j})|\right| + \sum_{j=J}^{\infty} 2 \cdot \frac{C|\omega|}{2^j} < \infty.$$

This shows that $\sum_{j=1}^{\infty} \ln|p(\frac{\omega}{2^j})|$ converges, completing the proof of the theorem. $\blacksquare$

For a finite sequence $p = \{p_k\}$, it can be shown, by using the Paley-Wiener theorem, that $\widehat{\phi}(\omega)$, defined by (10.1.3), is an entire function (i.e., a function analytic in the entire complex plane $\mathbb{C}$) and is the Fourier transform of some compactly supported distribution (also called generalized function) ϕ. Before showing that under a certain suitable condition imposed on the sequence p (formulated in terms of the transition operator T_p associated with p), the distribution ϕ is an L_2 function in Theorem 3 below, let us first prove that the support of the distribution ϕ is dictated by the support of p, when p is considered as a bi-infinite sequence by padding it with zeros .

Theorem 2 $\boxed{\text{Supports of refinable functions}}$ *For any finite sequence* $p = \{p_k\}_{k=0}^{N}$, *the solution* ϕ *of the refinement equation* (10.1.1) *is a distribution with* $supp(\phi) \subseteq [0, N]$.

Proof   From the refinement equation (10.1.1), it is clear that

$$supp(\phi) \subseteq \bigcup_{k \in [0,N]} supp(\phi(2 \cdot -k)) = \bigcup_{k \in [0,N]} \frac{1}{2}(supp(\phi) + k)$$

$$\subseteq \frac{1}{2}(supp(\phi) + [0, N]) = \frac{1}{2}[0, N] + \frac{1}{2}supp(\phi).$$

By applying this result to the second term on the right, we have

$$supp(\phi) \subseteq \frac{1}{2}[0, N] + \frac{1}{2}\left(\frac{1}{2}[0, N] + \frac{1}{2}supp(\phi)\right)$$

$$= \frac{1}{2}[0, N] + \frac{1}{2^2}[0, N] + \frac{1}{2^2}supp(\phi);$$

and more generally, by repeating the same argument, we obtain

$$\text{supp}(\phi) \subseteq \frac{1}{2}[0, N] + \frac{1}{2^2}[0, N] + \cdots + \frac{1}{2^j}[0, N] + \frac{1}{2^j}\text{supp}(\phi),$$

for $j \geq 1$. But $\lim_{j \to \infty} \frac{1}{2^j}\text{supp}(\phi) = 0$, since ϕ is compactly supported. Therefore, we may conclude that $\text{supp}(\phi) \subseteq \sum_{j=1}^{\infty} \frac{1}{2^j}[0, N] = [0, N]$, as desired. ∎

Remark 1 Following the proof of Theorem 2, one can show that if the refinement mask $p = \{p_k\}$ is supported on $[N_1, N_2]$, namely $p_k = 0$ for $k < N_1$ or $k > N_2$, then $\text{supp}(\phi) \subseteq [N_1, N_2]$ (see Exercise 2). ∎

Next, we provide a characterization for the normalized solution ϕ, as a function in $L_2(\mathbb{R})$, of the refinement equation, by imposing some restriction on the refinement mask. As remarked in Remark 1 of Sect. 9.1, in the rest of this chapter, a refinable function ϕ associated with $p = \{p_k\}$ means the normalized solution of the refinement equation with the condition $\widehat{\phi}(0) = 1$.

Theorem 3 | **Characterization of refinable functions in $L_2(\mathbb{R})$** | *Let $p = \{p_k\}_{k=0}^{N}$ and ϕ be the normalized solution of (10.1.1). Then the distribution ϕ is a function in $L_2(\mathbb{R})$, if and only if there exists some non-negative trigonometric polynomial $v(\omega) \in \mathbb{V}_{2N+1}$ with $v(0) > 0$, such that $v(\omega)$ is a 1-eigenfunction of the transition operator T_p.*

Proof Suppose $\phi \in L_2(\mathbb{R})$. Then it follows from Theorem 2 of Sect. 9.2 on p.448 that the Gramian function $G_\phi(\omega)$ of ϕ is in $\mathbb{V}_{2N+1}$ and $(T_p G_\phi)(\omega) = G_\phi(\omega)$. Clearly, $G_\phi(\omega) \geq 0$, and

$$G_\phi(0) = \sum_{k=-\infty}^{\infty} |\widehat{\phi}(2\pi k)|^2 \geq |\widehat{\phi}(0)|^2 = 1.$$

Thus, $v(\omega) = G_\phi(\omega)$ is a trigonometric polynomial that satisfies the required properties.

Conversely, suppose that there exists $v(\omega) \in \mathbb{V}_{2N+1}$ such that $(T_p v)(\omega) = v(\omega)$, $v(\omega) \geq 0$ and $v(0) > 0$. Then, by the Fundamental Lemma of transition operators in Sect. 9.1 of Chap. 9 (see (9.1.12) of Theorem 3 on p.444), we have

$$\int_{-2^n \pi}^{2^n \pi} \prod_{j=1}^{n} |p(\frac{\omega}{2^j})|^2 \, v(\frac{\omega}{2^n}) d\omega$$

$$= \int_{-\pi}^{\pi} (T_p^n v)(\omega) d\omega = \int_{-\pi}^{\pi} v(\omega) d\omega < \infty.$$

Thus, by applying Fatou's lemma, we may conclude that

$$v(0) \int_{\mathbb{R}} |\widehat{\phi}(\omega)|^2 d\omega = \int_{\mathbb{R}} |\widehat{\phi}(\omega)|^2 \, v(0) d\omega$$

$$= \int_{\mathbb{R}} \underline{\lim}_{n \to \infty} | \prod_{j=0}^{n} p\left(\frac{\omega}{2^j}\right) |^2 \, v\left(\frac{\omega}{2^n}\right) \chi_{2^n[-\pi,\pi]} d\omega$$

$$\leq \underline{\lim}_{n \to \infty} \int_{\mathbb{R}} | \prod_{j=0}^{n} p\left(\frac{\omega}{2^j}\right) |^2 \, v\left(\frac{\omega}{2^n}\right) \chi_{2^n[-\pi,\pi]} d\omega$$

$$= \underline{\lim}_{n \to \infty} \int_{-\pi}^{\pi} v(\omega) d\omega < \infty.$$

Hence, since $v(0) \neq 0$, we have

$$\int_{\mathbb{R}} |\widehat{\phi}(\omega)|^2 d\omega < \infty,$$

or $\phi \in L_2(\mathbb{R})$. ∎

Example 1 Apply Theorem 3 to show that the refinable function ϕ of the refinement equation

$$\phi(x) = \phi(2x) + \phi(2x - 2),$$

is a function in $L_2(\mathbb{R})$.

Solution First, we must find a 1-eigenfunction $v(\omega)$ of T_p that satisfies the conditions in Theorem 3. To do so, we compute the representation matrix $\mathcal{T}_p = \left[\frac{1}{2} a_{2j-k}\right]_{-2 \leq j,k \leq 2}$ of the transition operator T_p, by considering the Fourier series $\sum_j a_j e^{-ij\omega} = 4|p(\omega)|^2$ (refer to (9.1.8) on p.439), where

$$p(\omega) = \frac{1}{2}(1 + e^{-i2\omega}),$$

so that

$$\sum_j a_j e^{-ij\omega} = 4|p(\omega)|^2 = 2 + e^{-i2\omega} + e^{i2\omega}.$$

Therefore, the only nonzero coefficients a_j are given by

$$a_0 = 2, \; a_1 = a_{-1} = 1,$$

and this gives the representation matrix

$$\mathcal{T}_p = \left[\frac{1}{2} a_{2j-k}\right]_{-2 \le j,k \le 2} = \frac{1}{2} \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 2 & 0 & 1 & 0 & 0 \\ 1 & 0 & 2 & 0 & 1 \\ 0 & 0 & 1 & 0 & 2 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

It is not difficult to verify that $\mathbf{v}_0 = [0, 1, 2, 1, 0]^T$ is a (right) 1-eigenvector of $\mathcal{T}_p$. Thus,

$$v_0(\omega) = E(\omega)\mathbf{v}_0 = e^{i\omega} + 2 + e^{-i\omega} = 2 + 2\cos\omega$$

is a 1-eigenfunction of the transition operator T_p. Since the necessary conditions ($v_0(\omega) \ge 0$, for $\omega \in [-\pi, \pi]$; and $v_0(0) = 4 > 0$) are satisfied by the trigonometric polynomial $v_0(\omega) \in \mathbb{V}_5$, it follows from Theorem 3 that $\phi \in L_2(\mathbb{R})$. Indeed, it is easy to verify that the solution of the given refinement equation is the function $\phi(x) = \frac{1}{2}\chi_{[0,2)}(x)$.   ∎

Recall that a sequence $p = \{p_k\}_{k=0}^N$ is called a QMF if its two-scale symbol satisfies the identity

$$|p(\omega)|^2 + |p(\omega + \pi)|^2 = 1.$$

Therefore, for a QMF p, the constant trigonometric polynomial $1 \in \mathbb{V}_{2N+1}$ is an eigenfunction of the transition operator T_p corresponding to the eigenvalue 1, namely:

$$T_p 1 = |p(\omega/2)|^2 \cdot 1 + |p(\omega/2 + \pi)|^2 \cdot 1 = 1.$$

Since $v(\omega) = 1$ is non-negative and different from zero at $\omega = 0$, the necessary conditions in Theorem 3 are satisfied. This leads to the following result.

Corollary 1 $\boxed{\textbf{QMF implies } \phi \in L_2(\mathbb{R})}$ *Let N be any positive integer, $p = \{p_k\}_{k=0}^N$ any QMF, and ϕ the normalized solution of* (10.1.1). *Then the solution ϕ is a compactly supported function in $L_2(\mathbb{R})$, with* supp$(\phi) \subseteq [0, N]$.

Next, we will show that if the normalized solution of the refinement equation (10.1.2) is a function in $L_2(\mathbb{R})$, then this (scaling) function enjoys the property of partition of unity (PU). A compactly supported function $f(x)$ is said to have the PU, if it satisfies the identity

$$\sum_{\ell \in \mathbb{Z}} f(x - \ell) = 1, \ x \in \mathbb{R}. \tag{10.1.6}$$

Observe that since $f(x)$ is compactly supported, the infinite series in (10.1.6) is only a finite sum for each fixed value of x. Thus, the series always converges to a 1-periodic function.

In the next theorem we show that PU is equivalent to the moment condition (called the Strang-Fix condition), to be defined in (10.1.7) below.

Theorem 4 $\boxed{\textbf{PU} \Leftrightarrow \textbf{moment condition}}$ *A compactly supported function $f \in L_2(\mathbb{R})$ has the PU if and only if it satisfies the moment condition of order at least equal to 1; that is,*

$$\widehat{f}(0) = 1, \ \widehat{f}(2\pi k) = 0, \ k \in \mathbb{Z}\backslash\{0\}. \tag{10.1.7}$$

Proof Since f is compactly supported, $f \in L_2(\mathbb{R})$ implies $f \in L_1(\mathbb{R})$, and hence $g(x) = \sum_{\ell \in \mathbb{Z}} f(x - \ell)$ is in $L_1[0, 1]$. Observe that, for $k \in \mathbb{Z}$,

$$\int_0^1 g(x) e^{-i2\pi k x} dx = \int_0^1 \sum_{\ell \in \mathbb{Z}} f(x - \ell) e^{-i2\pi k x} dx$$

$$= \sum_{\ell \in \mathbb{Z}} \int_0^1 f(x - \ell) e^{-i2\pi k x} dx$$

$$= \sum_{\ell \in \mathbb{Z}} \int_{-\ell}^{1-\ell} f(u) e^{-i2\pi k u} du$$

$$= \int_{-\infty}^{\infty} f(u) e^{-i2\pi k u} du = \widehat{f}(2\pi k),$$

where the interchange of integration and summation in the second equality is valid, since $\sum_{\ell \in \mathbb{Z}}$ is a finite sum. Thus, f satisfies (10.1.6), if and only if it satisfies (10.1.7). ∎

Theorem 5 $\boxed{\phi \in L_2(\mathbb{R}) \textbf{ implies } \widehat{\phi}(2\pi k) = \delta_k}$ *Let ϕ be the normalized solution of the refinement equation (10.1.2), such that $\phi \in L_2(\mathbb{R})$. Then ϕ satisfies the moment condition of order at least equal to 1; that is,*

$$\widehat{\phi}(0) = 1, \ \widehat{\phi}(2\pi k) = 0, \ k \in \mathbb{Z}\backslash\{0\}. \tag{10.1.8}$$

Proof Since ϕ is compactly supported, $\phi \in L_2(\mathbb{R})$ implies $\phi \in L_1(\mathbb{R})$. Consequently, it follows from the Riemann-Lebesgue lemma that $\widehat{\phi}(\omega) \to 0$ as $\omega \to \infty$. Let $k \neq 0$. Then

$$\widehat{\phi}(2^n \cdot 2\pi k) = \prod_{j=1}^n p\left(\frac{2^n \cdot 2k\pi}{2^j}\right) \widehat{\phi}\left(\frac{2^n \cdot 2k\pi}{2^n}\right)$$

$$= \prod_{j=1}^n p(0) \widehat{\phi}(2k\pi) = \widehat{\phi}(2k\pi).$$

Hence, $\widehat{\phi}(2\pi k) = \widehat{\phi}(2^n \cdot 2\pi k) \to 0$ as $n \to \infty$; that is, $\widehat{\phi}(2\pi k) = 0$ for all integers $k \neq 0$, as desired. ∎

As an immediate consequence of Theorems 4 and 5, we have the following result.

Corollary 2 | $L_2(\mathbb{R})$ **for refinable** ϕ **implies** $\sum_k \phi(x-k) = 1$ | *A compactly sup-ported refinable function in $L_2(\mathbb{R})$ has the property of partition of unity.*

As another consequence of Theorem 5, observe that the Gramian function of a scaling function $\phi \in L_2(\mathbb{R})$ satisfies

$$G_\phi(0) = \sum_{k \in \mathbb{Z}} |\widehat{\phi}(2k\pi)|^2 = |\widehat{\phi}(0)|^2 = 1.$$

Also, recall from Definition 2 of Sect. 9.3 on p.456 that a trigonometric polynomial $p(\omega)$ is said to have sum-rule order $L \geq 1$, if L is the largest integer for which

$$p(0) = 1, \ \frac{d^\ell}{d\omega^\ell} p(\pi) = 0, \ \text{for } \ell = 0, 1, \ldots, L-1.$$

Theorem 6 | $G_\phi(\pi) > 0$ **implies** $p(\omega)$ **has at least first sum-rule order** | *Let $\phi \in L_2(\mathbb{R})$ be a compactly supported scaling function with refinement mask $\{p_k\}_{k=0}^N$. If $G_\phi(\pi) > 0$, then $p(\omega)$ has at least first sum-rule order; that is, $p(0) = 1$ and*

$$p(\pi) = 0. \tag{10.1.9}$$

Proof   To prove (10.1.9), we observe, from the fact that $T_p G_\phi = G_\phi$, that

$$G_\phi(0) = |p(0)|^2 G_\phi(0) + |p(\pi)|^2 G_\phi(\pi).$$

Thus, by the assumption that $p(0) = 1$ and the fact that $G_\phi(0) = 1$, we have $|p(\pi)|^2 G_\phi(\pi) = 0$, which implies $p(\pi) = 0$, as desired. ∎

It has been shown in Theorem 4 of Sect. 9.2, on p.455, that if ϕ is stable, then $G_\phi(\omega) > 0$. This fact and Theorem 6 lead to the following corollary.

Corollary 3 | **Stability of** ϕ **implies** $p(\omega)$ **has the sum-rule property** | *If a refinable function ϕ is stable, then its associated refinement mask $\{p_k\}_{k=0}^N$ has at least the first order sum-rule property.*

Example 2   Let $\phi(x) = \frac{1}{3}\chi_{[0,3)}(x)$ be the function considered in Example 4 of Sect. 9.2 in the previous chapter, on p.450, where it was shown that $G_\phi(\omega) = \frac{1}{3} + \frac{4}{9}\cos\omega + \frac{2}{9}\cos 2\omega$ and that $\phi(x)$ is not stable. However, it is easy to verify that $\phi(x)$ is a solution of the refinement equation

$$\phi(x) = \phi(2x) + \phi(2x-3),$$

and hence is a refinable function. Since

$$G_\phi(\pi) = \frac{1}{3} + \frac{4}{9}\cos\pi + \frac{2}{9}\cos 2\pi = \frac{1}{9} > 0,$$

it follows from Theorem 6 that the refinement mask p of ϕ should have at least the first sum-rule order . Indeed, this is true, as

$$p(\omega) = \frac{1}{2}(1 + e^{-i3\omega}),$$

which implies that $p(0) = 1$ and $p(\pi) = \frac{1}{2}(1 - 1) = 0$.   ■

Example 3 Let $\phi(x) = \frac{1}{2}\chi_{[0,2)}(x)$ be the refinable function considered in Example 1 of this section. The associated two-scale symbol $p(\omega)$ is given by

$$p(\omega) = \frac{1}{2}(1 + e^{-i2\omega}).$$

Although $\phi \in L_2(\mathbb{R})$, the refinement mask p does not satisfy the sum-rule condition, since $p(\pi) = \frac{1}{2}(1 + 1) = 1 \neq 0$. Observe that this does not contradict Theorem 6, because the refinable function ϕ does not satisfy the condition $G_\phi(\pi) > 0$. Observe that

$$G_\phi(\omega) = \frac{1}{2} + \frac{1}{2}\cos\omega,$$

and $G_\phi(\pi) = 0$ (see Exercise 3).   ■

<div align="center">

Excercises

</div>

Exercise 1 Show that $-2x \leq \ln(1 - x)$ for $0 \leq x \leq \frac{1}{2}$.

Exercise 2 Let ϕ be the refinable distribution associated with $p = \{p_k\}_{k=N_1}^{N_2}$. Show that $\mathrm{supp}(\phi) \subseteq [N_1, N_2]$.

Exercise 3 Let $\phi(x) = \frac{1}{2}\chi_{[0,2)}(x)$ be the function considered in Example 3. Show that $G_\phi(\omega) = \frac{1}{2} + \frac{1}{2}\cos\omega$.

Exercise 4 Let $\phi(x) = \frac{1}{3}\chi_{[0,3)}(x)$ be the refinable function considered in Example 2. Verify that it satisfies the refinement equation $\phi(x) = \phi(2x) + \phi(2x - 3)$. Then compute the representation matrix of the transition operator T_p associated with the refinement mask. Next obtain the 1-eigenfunction $v(\omega)$ of T_p. Finally, verify that $v(\omega)$ satisfies the necessary conditions in Theorem 3.

Exercise 5 Let $\phi(x) = \frac{1}{3}\chi_{[0,3)}(x)$ be a refinable function. Show that its two-scale symbol is a QMF, and hence deduce from Corollary 1 that $\phi \in L_2(\mathbb{R})$.

Exercise 6 Let $p(\omega) = \frac{1}{2}(1 + e^{-i2\omega})$ be the symbol of $\phi(x) = \frac{1}{2}\chi_{[0,2)}(x)$, considered in Example 3. Find the eigenvalues of T_p and 1-eigenfunctions of T_p.

Exercise 7 Let $p = \{p_k\}_{k=0}^3$ be a refinement mask with

$$p_0 = p_3 = \frac{1}{4}, \; p_1 = p_2 = \frac{3}{4}.$$

Find the 1-eigenfunctions of T_p, and then apply Theorem 3 to decide whether or not the corresponding refinable distribution ϕ is a function in $L_2(\mathbb{R})$. (For reference, see Exercise 4 on p.444.)

Exercise 8 Repeat Exercise 7 for $p = \{p_k\}_{k=0}^3$ with

$$p_0 = p_1 = p_2 = p_3 = \frac{1}{2}.$$

10.2 Stability and Orthogonality of Refinable Functions

In Sect. 9.2 of Chap. 9, we have presented the characterizations of the properties of orthogonality and stability of the refinable function ϕ in terms of the Gramian function $G_\phi(\omega)$ of ϕ (see Theorem 3 on p.449 and Theorem 4 on p.449). These characterizations require the explicit expression of ϕ or its Fourier transform for the formulation and calculation of the Gramian function $G_\phi(\omega)$. However, in general, the existence of a refinable function ϕ is only in terms of some infinite product, as studied in the previous section, and there is no explicit expression of ϕ or its Fourier transform. Such examples include all compactly supported orthogonal scaling functions that are continuous, as studied in both Chaps. 8 and 9. In this situation, we must find some other way to characterize orthogonality and stability. In this section, we provide certain characterizations in terms of the refinement mask p and its transition operator T_p.

Firstly, let us recall the notion of Condition E on a matrix (or a linear operator on a finite-dimensional vector space), as introduced in Definition 1 of Sect. 9.3 of Chap. 9, on p.455.

$\boxed{\textbf{Condition E}}$ A matrix (or a linear operator T on a finite-dimensional space) is said to satisfy **Condition E**, to be denoted by $T \in$ E, if 1 is a simple eigenvalue of T and any other eigenvalue λ of T satisfies $|\lambda| < 1$. ∎

Theorem 1 $\boxed{\textbf{Characterization for stability of refinable } \phi}$ *Let ϕ be a compactly supported refinable function with refinement mask $p = \{p_k\}_{k=0}^N$. Then ϕ is stable if and only if the following two conditions are satisfied*:

(i) $p(\pi) = 0$; *and*
(ii) $T_p \in$ E, *and there exists a 1-eigenfunction $v_0(\omega)$ of T_p that satisfies either* $v_0(\omega) > 0$, *or $v_0(\omega) < 0$, for all $\omega \in \mathbb{R}$.*

The proof of Theorem 1 will be provided at the end of this section.

Example 1 Let ϕ be the refinable function with refinement mask $p = \{p_k\}_{k=0}^2$, where

$$p_0 = \frac{3}{4}, \; p_1 = 1, \; p_2 = \frac{1}{4}.$$

Apply Theorem 1 to decide whether or not ϕ is stable.

Solution   The nonzero values of a_j can be computed from the formula $a_j = \sum_{s=0}^{N} p_s p_{s-j}$, namely:

$$a_0 = \frac{13}{8}, \; a_1 = a_{-1} = 2, \; a_2 = a_{-2} = \frac{3}{16}.$$

Thus, we have the representation matrix

$$\mathcal{T}_p = \left[\frac{1}{2} a_{2j-k} \right]_{-2 \leq j,k \leq 2} = \frac{1}{32} \begin{bmatrix} 3 & 0 & 0 & 0 & 0 \\ 26 & 16 & 3 & 0 & 0 \\ 3 & 16 & 26 & 16 & 3 \\ 0 & 0 & 3 & 16 & 26 \\ 0 & 0 & 0 & 0 & 3 \end{bmatrix}$$

of the transition operator $\mathcal{T}_p$. To compute the 1-eigenvector of $\mathcal{T}_p$, we first observe that each of the first and last rows of the matrix $\mathcal{T}_p$ has only one nonzero entry, so that the determinant of the matrix $\lambda I_5 - \mathcal{T}_p$ of dimension 5 can be simplified to be

$$\det(\lambda I_5 - \mathcal{T}_p) = (\lambda - \frac{3}{32})^2 \begin{vmatrix} \lambda - \frac{1}{2} & -\frac{3}{32} & 0 \\ -\frac{1}{2} & \lambda - \frac{13}{16} & -\frac{1}{2} \\ 0 & -\frac{3}{32} & \lambda - \frac{1}{2} \end{vmatrix}$$

$$= (\lambda - \frac{3}{32})^2 \begin{vmatrix} \lambda - 1 & \lambda - 1 & \lambda - 1 \\ -\frac{1}{2} & \lambda - \frac{13}{16} & -\frac{1}{2} \\ 0 & -\frac{3}{32} & \lambda - \frac{1}{2} \end{vmatrix}$$

(by adding row-2 and row-3 to row-1)

$$= (\lambda - \frac{3}{32})^2 \begin{vmatrix} \lambda - 1 & 0 & 0 \\ -\frac{1}{2} & \lambda - \frac{13}{16} + \frac{1}{2} & 0 \\ 0 & -\frac{3}{32} & \lambda - \frac{1}{2} \end{vmatrix}$$

(by subtracting col-1 from col-2 and col-3)

$$= (\lambda - \frac{3}{32})^2 (\lambda - 1)(\lambda - \frac{1}{2})(\lambda - \frac{5}{16}).$$

Thus, the eigenvalues of $\mathcal{T}_p$ are given by

$$1, \frac{1}{2}, \frac{5}{16}, \frac{3}{32}, \frac{3}{32}.$$

This ensures that $T_p \in E$. In addition, it is easy to verify that the vector

$$[0, 3, 16, 3, 0]$$

is a 1-eigenvector of $\mathcal{T}_p$, so that

$$v(\omega) = 3e^{i\omega} + 16 + 3e^{-i\omega} = 16 + 6\cos\omega$$

is a 1-eigenfunction of T_p. Clearly, $v(\omega) > 0$ for any $\omega \in \mathbb{R}$. Furthermore, p has at least sum-rule order 1. Thus, by Theorem 1, we may conclude that ϕ is stable.   ∎

Next, we give a characterization on the orthogonality of compactly supported refinable functions.

Theorem 2 | **Characterization of orthogonal refinable functions** | *Let ϕ be a compactly supported refinable function with refinement mask $p = \{p_k\}_{k=0}^{N}$. Then ϕ is in $L_2(\mathbb{R})$ and orthogonal, if and only if the following conditions on p and the transition operator T_p are satisfied:*

  (i) *p is a QMF;*
 (ii) *$p(\pi) = 0$; and*
(iii) *$T_p \in E$.*

Proof Suppose $\phi \in L_2(\mathbb{R})$ is orthogonal. Then statements (ii) and (iii) follow from Theorem 1, since orthogonality implies stability of ϕ. Although it was already shown in Theorem 2 on p.453 that if ϕ is orthogonal, then the corresponding refinement mask p is a QMF, we re-write the proof by only considering the transition operator. Indeed, by the orthogonality of ϕ, it follows that $G_\phi(\omega) = 1$ (see Theorem 3 on p.449). Thus, $T_p 1 = 1$; that is,

$$|p(\omega/2)|^2 + |p(\omega/2 + \pi)|^2 = 1,$$

so that p is a QMF.

Conversely, suppose that (i)–(iii) in Theorem 2 hold. Then, since p is a QMF, it follows from Corollary 1 on p.506 that $\phi \in L_2(\mathbb{R})$. Hence, G_ϕ is a 1-eigenfunction of T_p. Also, since $T_p \in E$ and 1 is a 1-eigenfunction of T_p, we may conclude that $G_\phi(\omega) = c_0 \cdot 1 = c_0$, for some constant c_0. On the other hand, by Theorem 5 of the previous section on p.507, $\widehat{\phi}(2\pi k) = \delta_k$, which implies that $G_\phi(0) = 1$, so that $c_0 = 1$. Hence, $G_\phi(\omega) = 1$, which is equivalent to the statement that ϕ is orthogonal.   ∎

Remark 1 Let p be a finite QMF that has at least sum-rule order 1, and ϕ be the refinable function with refinement mask p. Then it follows from Theorem 2 that

$$\phi \text{ is orthogonal} \iff T_p \in \mathrm{E}.$$

∎

Example 2 Let $p = \{p_k\}_{k=0}^3$ with

$$p_0 = \frac{1 - \sqrt{3}}{4}, \ p_1 = \frac{3 - \sqrt{3}}{4}, \ p_2 = \frac{3 + \sqrt{3}}{4}, \ p_3 = \frac{1 + \sqrt{3}}{4},$$

be the refinement mask for the scaling function ϕ, associated with the D_4 wavelet, studied in Sect. 9.3 of the Chap. 9 (see (9.3.21) on p.460). Apply Theorem 2 to verify that ϕ is orthogonal.

Solution Since $p(\omega)$ is a QMF and has sum-rule property of order ≥ 1, to prove that ϕ is orthogonal, it is sufficient to show that T_p satisfies Condition E. By Remark 5 on p.443, the verification is in turn reduced to showing that $T_p = [\frac{1}{2} a_{2j-k}] \in \mathrm{E}$.

To compute a_j from the Fourier series representation $\sum_j a_j e^{-ij\omega} = 4|p(\omega)|^2$ (see (9.1.8) on p.439), recall from the construction of the D_4 filter on p.459, that

$$
\begin{aligned}
|p(\omega)|^2 &= \cos^4 \frac{\omega}{2} \left(1 + 2 \sin^2 \frac{\omega}{2}\right) \\
&= \frac{1}{16} \left(2 + e^{i\omega} + e^{-i\omega}\right)^2 \left(2 - \frac{1}{2} e^{i\omega} - \frac{1}{2} e^{-i\omega}\right) \\
&= \frac{1}{32} \left(16 + 9 e^{i\omega} + 9 e^{-i\omega} - e^{i3\omega} - e^{-i3\omega}\right).
\end{aligned}
$$

Thus, the nonzero a_k are given by

$$a_0 = 2, \ a_1 = a_{-1} = \frac{9}{8}, \ a_3 = a_{-3} = -\frac{1}{8},$$

so that the representation matrix of the transition operator T_p is given by

$$
T_p = \left[\frac{1}{2} a_{2j-k}\right]_{-3 \leq j,k \leq 3} = \frac{1}{16}
\begin{bmatrix}
-1 & 0 & 0 & 0 & 0 & 0 & 0 \\
9 & 0 & -1 & 0 & 0 & 0 & 0 \\
9 & 16 & 9 & 0 & -1 & 0 & 0 \\
-1 & 0 & 9 & 16 & 9 & 0 & -1 \\
0 & 0 & -1 & 0 & 9 & 16 & 9 \\
0 & 0 & 0 & 0 & -1 & 0 & 9 \\
0 & 0 & 0 & 0 & 0 & 0 & -1
\end{bmatrix}.
$$

To compute the eigenvalues of T_p, observe that each of the first and last rows, as well as the 4th column, has one nonzero entry. Thus, the desired determinant can be reduced to

$$
\det(\lambda I_7 - \mathcal{T}_p) = (\lambda + \frac{1}{16})^2(\lambda - 1)
\begin{vmatrix}
\lambda & \frac{1}{16} & 0 & 0 \\
-1 & \lambda - \frac{9}{16} & \frac{1}{16} & 0 \\
0 & \frac{1}{16} & \lambda - \frac{9}{16} & -1 \\
0 & 0 & \frac{1}{16} & \lambda
\end{vmatrix}
$$

$$
= (\lambda + \frac{1}{16})^2(\lambda - 1)(\lambda - \frac{1}{2})(\lambda - \frac{1}{4})^2(\lambda - \frac{1}{8}),
$$

where the calculation details for the last equality are omitted. Thus, the eigenvalues of $\mathcal{T}_p$ are

$$
1, \ \frac{1}{2}, \ \frac{1}{4}, \ \frac{1}{4}, \ \frac{1}{8}, \ -\frac{1}{16}, \ -\frac{1}{16},
$$

implying that $\mathcal{T}_p \in$ E. Therefore, the scaling function ϕ is orthogonal.

From $\mathcal{T}_p$, it is easy to verify that the vector

$$
[0, 0, 0, 1, 0, 0, 0]
$$

is an eigenvector of $\mathcal{T}_p$ corresponding to the eigenvalue 1, so that the constant trigonometric polynomial $v(\omega) = 1$ is a 1-eigenfunction of $\mathcal{T}_p$. Since $G_\phi(\omega)$ is also a 1-eigenfunction of $\mathcal{T}_p$, we may conclude that $G_\phi(\omega) = cv(\omega) = c$ for some constant c. But from $G_\phi(0) = 1$, we have $c = 1$. Thus, $G_\phi(\omega) = 1$ for all $\omega \in \mathbb{R}$. This is exactly the orthogonality condition on ϕ given in Theorem 3 on p.449.   ∎

We end this section by giving the proof of Theorem 1.

Proof of Theorem 1   First, suppose that (i) and (ii) in Theorem 1 hold. Then since $v_0(0) \neq 0$, it follows from Theorem 3 that $\phi \in L_2(\mathbb{R})$, so that $G_\phi(\omega) \in \mathbb{V}_{2N+1}$ and $T_p G_\phi = G_\phi$. Now, because 1 is a simple eigenvalue of T_p, $G_\phi(\omega) = c_0 v_0(\omega)$ for some constant $c_0 \neq 0$. Thus, since $G_\phi(\omega) \geq 0$ and $v_0(\omega)$ is never zero on $\mathbb{R}$, we may conclude that $G_\phi(\omega) > 0$ for all $\omega \in \mathbb{R}$. Therefore, by Theorem 4 on p.449, ϕ is stable.

Conversely, assume that ϕ is stable. Then statement (i) follows from Corollary 3 in Sect. 10.1 (see p.508). In addition, the stability of ϕ implies that $G_\phi \in \mathbb{V}_{2N+1}$ and $T_p G_\phi = G_\phi$ and $c \leq G_\phi(\omega) \leq C, \omega \in \mathbb{R}$, for some constants $c, C > 0$. Thus, $v_0(\omega) = G_\phi(\omega)$ is a 1-eigenfunction of T_p which is never zero. Therefore, to complete the proof of Theorem 1, it is sufficient to show that $T_p \in$ E.

For any $f, g \in \mathbb{V}_{2N+1}$, we have, by Theorem 3 in Sect. 9.1 on p.443,

$$
\int_{-\pi}^{\pi} g(\omega)(T_p^n f)(\omega) d\omega = \int_{-2^n \pi}^{2^n \pi} g(\omega) \prod_{j=1}^{n} |p(\frac{\omega}{2^j})|^2 f(\frac{\omega}{2^n}) d\omega
$$

$$
= \int_{-2^n \pi}^{2^n \pi} g(\omega) \prod_{j=1}^{n} |p(\frac{\omega}{2^j})|^2 f(\frac{\omega}{2^n}) \sum_{k \in \mathbb{Z}} |\widehat{\phi}(\frac{\omega}{2^n} + 2k\pi)|^2 / G_\phi(\frac{\omega}{2^n}) d\omega
$$

$$
= \sum_{k \in \mathbb{Z}} \int_{-2^n \pi + 2k\pi 2^n}^{2^n \pi + 2k\pi 2^n} g(u) \prod_{j=1}^{n} |p(\frac{u}{2^j})|^2 |\widehat{\phi}(\frac{u}{2^n})|^2 f(\frac{u}{2^n}) / G_\phi(\frac{u}{2^n}) du
$$

$$= \int_{\mathbb{R}} g(u) \prod_{j=1}^{n} |p(\frac{u}{2^j})|^2 |\widehat{\phi}(\frac{u}{2^n})|^2 f(\frac{u}{2^n})/G_\phi(\frac{u}{2^n}) du$$

$$= \int_{\mathbb{R}} g(u) |\widehat{\phi}(u)|^2 f(\frac{u}{2^n})/G_\phi(\frac{u}{2^n}) du,$$

where the third equality is obtained by substituting $u = \omega + 2k\pi 2^n$ and observing the 2π-periodicity of g, f, G_ϕ; while the last equality follows from the refinement equation for ϕ; that is, $\widehat{\phi}(\omega) = p(\frac{\omega}{2})\widehat{\phi}(\frac{\omega}{2})$. Thus,

$$\int_{-\pi}^{\pi} g(\omega)(T_p^n f)(\omega)d\omega = \int_{\mathbb{R}} g(\omega)|\widehat{\phi}(\omega)|^2 f(\frac{\omega}{2^n})/G_\phi(\frac{\omega}{2^n})d\omega. \qquad (10.2.1)$$

Now let λ be an eigenvalue of T_p with eigenfunction $v(\omega) \in \mathbb{V}_{2N+1}$. Then, by (10.2.1), we have

$$\lambda^n \int_{-\pi}^{\pi} \overline{v(\omega)} v(\omega)d\omega = \int_{-\pi}^{\pi} \overline{v(\omega)}(T_p^n v)(\omega)d\omega$$

$$= \int_{\mathbb{R}} \overline{v(\omega)} |\widehat{\phi}(\omega)|^2 v(\frac{\omega}{2^n})/G_\phi(\frac{\omega}{2^n})d\omega,$$

where the integrand of the last integral satisfies

$$\left| \overline{v(\omega)} |\widehat{\phi}(\omega)|^2 v(\frac{\omega}{2^n})/G_\phi(\frac{\omega}{2^n}) \right| \le C_1 |\widehat{\phi}(\omega)|^2, \ \omega \in \mathbb{R},$$

for some positive constant C_1, since a trigonometric polynomial $v(\omega)$ is bounded and $|1/G_\phi(\frac{\omega}{2^n})| \le \frac{1}{c}$. Thus, by Lebesgue's dominated convergence theorem, we have

$$\lim_{n\to\infty} \lambda^n = \lim_{n\to\infty} \frac{1}{\int_{-\pi}^{\pi} |v(\omega)|^2 d\omega} \int_{\mathbb{R}} \overline{v(\omega)} |\widehat{\phi}(\omega)|^2 v(\frac{\omega}{2^n})/G_\phi(\frac{\omega}{2^n})d\omega$$

$$= \frac{1}{\int_{-\pi}^{\pi} |v(\omega)|^2 d\omega} \int_{\mathbb{R}} \overline{v(\omega)} |\widehat{\phi}(\omega)|^2 v(0)/G_\phi(0)d\omega < \infty.$$

Therefore, $\lim_{n\to\infty} \lambda^n$ exists, which implies that we have either $\lambda = 1$ or $|\lambda| < 1$. Hence, to complete the argument in showing $T_p \in E$, we only need to show that 1 is a simple eigenvalue.

First, we know that 1 is an eigenvalue of T_p, since $T_p G_\phi(\omega) = G_\phi(\omega)$ and $G_\phi(\omega) \ne 0$. Next, let $v(\omega) \in \mathbb{V}_{2N+1}$ be an arbitrary 1-eigenfunction of T_p, and let

$$f(\omega) = v(\omega) - v(0)G_\phi(\omega).$$

Then $(T_p f)(\omega) = f(\omega)$, and $f(0) = 0$, since $G_\phi(0) = 1$. Hence, by (10.2.1) again,

$$\int_{-\pi}^{\pi} |f(\omega)|^2 d\omega = \int_{-\pi}^{\pi} \overline{f(\omega)}(T_p^n f)\omega)d\omega$$

$$= \int_{\mathbb{R}} \overline{f(\omega)} |\widehat{\phi}(\omega)|^2 f(\frac{\omega}{2^n})\Big)/G_\phi(\frac{\omega}{2^n})d\omega$$

$$\rightarrow \int_{\mathbb{R}} \overline{f(\omega)} |\widehat{\phi}(\omega)|^2 f(0)/G_\phi(0)d\omega = 0,$$

as $n \rightarrow \infty$. Thus $f = 0$; that is, $v(\omega) = v(0)G_\phi(\omega)$. Therefore, up to a constant, $G_\phi(\omega)$ is the unique 1-eigenfunction of T_p. This means that the geometric multiplicity of the eigenvalue 1 is 1.

Finally, let us show that 1 is nondegenerate. Assume to the contrary that it is not the case. Then there exists $v(\omega) \in \mathbb{V}_{2N+1}$ such that $T_p v(\omega) = G_\phi(\omega) + v(\omega)$. Again, let $f(\omega) = v(\omega) - v(0)G_\phi(\omega)$. Then $f(0) = 0$ and, by (10.2.1), we have

$$\int_{-\pi}^{\pi} T_p^n f(\omega)d\omega = \int_{\mathbb{R}} |\widehat{\phi}(\omega)|^2 f(\frac{\omega}{2^n})/G_\phi(\frac{\omega}{2^n})d\omega$$

$$\rightarrow \int_{\mathbb{R}} |\widehat{\phi}(\omega)|^2 f(0)/G_\phi(0)d\omega = 0,$$

by allowing $n \rightarrow \infty$. On the other hand, observe that

$$T_p^n f(\omega) = T_p^n v(\omega) - v(0)G_\phi(\omega) = nG_\phi(\omega) + v(\omega) - v(0)G_\phi(\omega),$$

so that

$$\int_{-\pi}^{\pi} T_p^n f(\omega)d\omega = \int_{-\pi}^{\pi} nG_\phi(\omega) + v(\omega) - v(0)G_\phi(\omega)d\omega \rightarrow \infty,$$

as $n \rightarrow \infty$. This is a contradiction to the fact that $\int_{-\pi}^{\pi} T_p^n f(\omega)d\omega \rightarrow 0$ as $n \rightarrow \infty$. Thus, 1 is indeed a simple eigenvalue of T_p, and hence, $T_p \in E$, as desired.   ∎

Exercises

Exercise 1 Let $p = \{p_k\}_{k=0}^2$ with

$$p_0 = \frac{1}{4}, \ p_1 = 1, \ p_2 = \frac{3}{4},$$

be a refinement mask. Apply Theorem 1 to decide whether or not the associated refinable function ϕ is stable.

Exercise 2 Repeat Exercise 1 for $p = \{p_k\}_{k=0}^2$ with

$$p_0 = \frac{2}{3}, \; p_1 = 1, \; p_2 = \frac{1}{3}.$$

Exercise 3 Repeat Exercise 1 for $p = \{p_k\}_{k=0}^3$ with

$$p_0 = p_3 = \frac{1}{4}, \; p_1 = p_2 = \frac{3}{4}.$$

(For reference, see Exercise 4 on p.444 and Exercise 7 on p.510.)

Exercise 4 Repeat Exercise 1 for $p = \{p_k\}_{k=0}^3$ with

$$p_0 = p_1 = p_2 = p_3 = \frac{1}{2}.$$

(For reference, see Exercise 8 on p.510.)

Exercise 5 Repeat Exercise 1 for $p = \{p_k\}_{k=0}^3$ with

$$p_0 = 1, \; p_1 = \frac{1}{2}, \; p_2 = 0, \; p_3 = \frac{1}{2}.$$

Exercise 6 Let $p = \{p_k\}_{k=0}^5$ be the refinement mask for the refinable function ϕ associated with the D_6 wavelet. By following Example 2, compute the eigenvalues of the associated transition operator T_p, and then conclude that ϕ is orthogonal.

10.3 Cascade Algorithms

In general, compactly supported scaling functions (and their corresponding wavelets) are difficult to construct. In fact, with a few exceptions, such as Cardinal B-splines, they are only defined by the inverse Fourier transform of certain infinite products, as in (10.1.3) of Sect. 10.1, where the trigonometric polynomial $p(\omega)$ is the two-scale symbol of a given finite refinement mask p. Hence, such scaling functions do not have explicit expressions, and the plotting of their graphs can only depend on the sequences p. For convenience, we will assume, without loss of generality, in this section that the support of $p = \{p_k\}_{k=0}^N$ is $[0, N]$, meaning that both p_0 and p_N are nonzero. It was proved in Sect. 10.1 that if the refinement equation has a solution ϕ which is a function, then the support of this function is within $[0, N]$ (see Theorem 2 in Sect. 10.1). Hence, to plot the graph $y = \phi(x)$, only the values of $\phi(x)$, for $0 \le x \le N$, are needed. Probably the most obvious way to render this graph is to apply the refinement equation directly to compute the values of $\phi(x)$ at the dyadic points $x = k/2^j$ for $k = 0, \dots, N2^j$. To do so, the first step is to compute the values of $\phi(x)$ at the integers; namely, the components of the vector

$$\mathbf{y}_0 = [\phi(0), \phi(1), \ldots, \phi(N)]^T.$$

It is easy to show that this vector $\mathbf{y}_0$ is a 1-eigenvector of the $(N+1)$-dimensional square matrix

$$\mathcal{P}_{N+1} = \Big[p_{2j-k} \Big]_{0 \leq j,k \leq N}, \tag{10.3.1}$$

with the (j, k)th entry being p_{2j-k}, where it is understood that $p_k = 0$ for $k < 0$ or $k > N$. If the values $\phi(0), \phi(1/2^{j-1}), \ldots, \phi(2^{j-1}N/2^{j-1})$ have been computed, then the values $\phi(m/2^j)$ can be found by computing the sum

$$\phi(m/2^j) = \sum_{k=0}^{N} p_k \phi(m/2^{j-1} - k) = \sum_{k=0}^{N} p_k \phi((m - 2^{j-1}k)/2^{j-1}), \tag{10.3.2}$$

for each $m = 1, \ldots, 2^j N - 1$. To find the values of $\phi(x)$ at $x = m/2^n$ for any desired positive integer n, this computational process is repeated iteratively, for $j = 1, \ldots, n$ (see Exercises 1, 2, and 3). Of course, this process requires that the refinement mask p must be so chosen that 1 is an eigenvalue of the matrix $\mathcal{P}_{N+1}$ in (10.3.1). In this regard, we note that when p has at least sum-rule order 1 (this is a minimal condition for the existence of a scaling function), p satisfies

$$\sum_k p_{2k} = \sum_k p_{2k+1} = 1,$$

which implies that the $1 \times (N+1)$ row vector

$$\mathbf{v}_0 = [1, 1, \ldots, 1]$$

satisfies $\mathbf{v}_0 \mathcal{P}_{N+1} = \mathbf{v}_0$. Thus, in this case, 1 is indeed an eigenvalue of $\mathcal{P}_{N+1}$. If 1 is a simple eigenvalue of $\mathcal{P}_{N+1}$, then $\mathbf{y}_0 = [\phi(0), \phi(1), \ldots, \phi(N)]^T$ is determined as the unique (up to a constant) (right) 1-eigenvector of $\mathcal{P}_{N+1}$.

In this section, we will introduce and study another approach to render $\phi(x)$, but again by using the refinement mask p. Instead of evaluating ϕ on the dense set of dyadic points in the interval $(0, N)$, we approximate $\phi(x)$ for all $x \in [0, N]$. The motivation to this approach is to observe that if a sequence of functions ϕ_n converges to ϕ, then ϕ_{n-1} also converges to ϕ, as n tends to infinity. Hence, if we consider the functional equations

$$\phi_n(x) = \sum_{k=0}^{N} p_k \phi_{n-1}(2x - k),$$

for each $n = 1, 2, \ldots$, the limit function $\phi(x)$ satisfies the refinement equation, and hence is the refinable function we wish to find. This is also an iterative scheme, but with some initial function ϕ_0, as opposed to using the components of the 1-eigenvector $\mathbf{y}_0$ of the matrix $\mathcal{P}_{N+1}$ as the initial values in computing the values in

(10.3.2) iteratively. To realize this approximation approach rigorously, we introduce the **refinement operator** Q_p, defined by

$$Q_p f(x) = \sum_{k \in \mathbb{Z}} p_k f(2x - k), \qquad f \in L_2(\mathbb{R}), \tag{10.3.3}$$

and for $n = 1, 2, \ldots$, set

$$Q^{n+1} = Q^n Q, \ Q^1 = Q,$$

so that for any initial function $\phi_0 \in L_2(\mathbb{R})$, we may also use the notation:

$$\boxed{\phi_n = Q_p^n \phi_0 = Q_p^{n-1}(Q_p \phi_0)}.$$

In the above discussion and throughout this section, the refinement mask $p = \{p_k\}_{k=0}^N$ is always assumed to satisfy $\sum_{k=0}^N p_k = 2$, with nonzero p_0 and p_N. The iterative process of computing $\phi_1, \phi_2, \ldots$, with a compactly supported $L_2(\mathbb{R})$ function ϕ_0 as initial function, is called the **cascade algorithm**.

We remark that we have abused the use of notations: while the notation $\phi_n = Q_p^n \phi_0$ is used in this section to denote the result obtained after n iterative steps of the cascade algorithm, it was used in other sections to denote $\phi(x - n)$, which is the shift of $\phi(x)$ by n. Returning to the notation for the cascade algorithm, we observe that

$$\widehat{\phi}_n(\omega) = \prod_{j=1}^n p(\frac{\omega}{2^j}) \, \widehat{\phi}_0(\frac{\omega}{2^n}).$$

Thus, according to Theorem 1 of Sect. 10.1, on p.502, if $\widehat{\phi}_0(0) = 1$, then $\widehat{\phi}_n(\omega)$ converges pointwise to $\widehat{\phi}(\omega)$ for all $\omega \in \mathbb{R}$.

Example 1 Let $p = \{p_k\}_{k=0}^2$ be the refinement mask of the hat function, with $p_0 = p_2 = \frac{1}{2}$, $p_1 = 1$. With $\phi_0(x) = \chi_{[0,1)}(x)$, we have

$$\phi_1(x) = (Q_p \phi_0)(x) = \frac{1}{2}\phi_0(2x) + \phi_0(2x - 1) + \frac{1}{2}\phi_0(2x - 2)$$

$$= \begin{cases} \frac{1}{2}, & \text{for } 0 \leq x < 0.5, \\ 1, & \text{for } 0.5 \leq x < 1, \\ \frac{1}{2}, & \text{for } 0 \leq x < 1.5, \\ 0, & \text{elsewhere.} \end{cases}$$

Then, from ϕ_1, we obtain ϕ_2, as given by
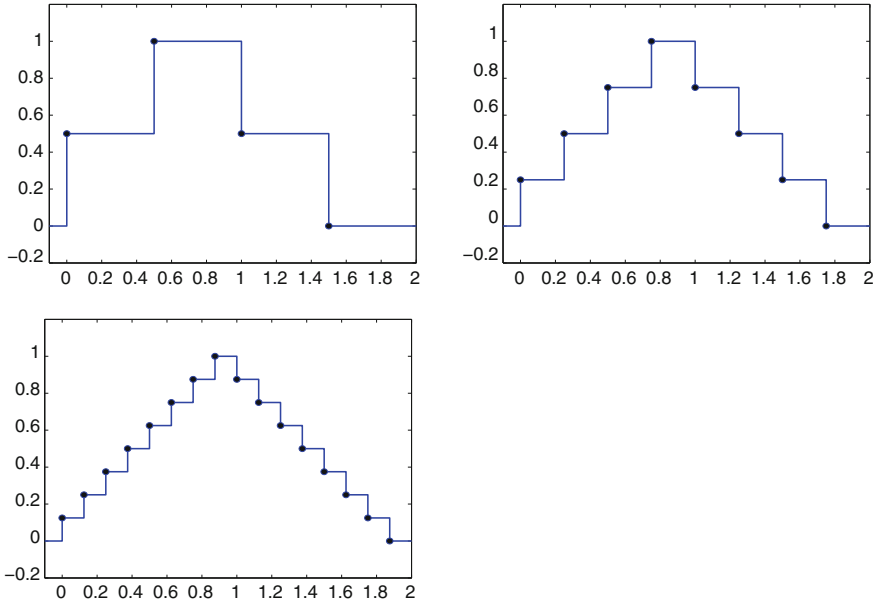
Fig. 10.1. $\phi_1 = Q_p\phi_0$, $\phi_2 = Q_p^2\phi_0$, $\phi_3 = Q_p^3\phi_0$ obtained by cascade algorithm with refinement mask of hat function and $\phi_0 = \chi_{[0,1)}(x)$

$$\phi_2(x) = Q_p\phi_1(x) = \frac{1}{2}\phi_1(2x) + \phi_1(2x-1) + \frac{1}{2}\phi_1(2x-2)$$

$$= \begin{cases} \frac{1}{4}, & \text{for } x \in [0, 0.25) \cup [1.5, 1.75), \\ \frac{1}{2}, & \text{for } x \in [0.25, 0.5) \cup [1.25, 1.5), \\ \frac{3}{4}, & \text{for } x \in [0.5, 0.75) \cup [1, 1.25), \\ 1, & \text{for } x \in [0.75, 1), \\ 0, & \text{elsewhere.} \end{cases}$$

In this way, we can compute ϕ_n for $n \geq 3$. The graphs of ϕ_1, ϕ_2, ϕ_3 are displayed in Fig. 10.1. From the graphical results, it appears that the sequence ϕ_n converges in $L_2(\mathbb{R})$ to the hat function ϕ. ∎

Example 2 Let $p = \{p_k\}_{k=0}^2$ with $p_0 = p_2 = 1$, $p_1 = 0$ be the refinement mask of $\phi(x) = \frac{1}{2}\chi_{[0,2)}(x)$, considered in Example 1 on p.505 and Example 3 on p.509. By applying the cascade algorithm associated with this mask with initial function $\phi_0(x) = \chi_{[0,1)}(x)$, we obtain

$$\phi_1(x) = Q_p\phi_0(x) = \phi_0(2x) + 0\phi_0(2x-1) + \phi_0(2x-2)$$

$$= \begin{cases} 1, & \text{for } x \in [0, 0.5) \cup [1, 1.5), \\ 0, & \text{elsewhere.} \end{cases}$$
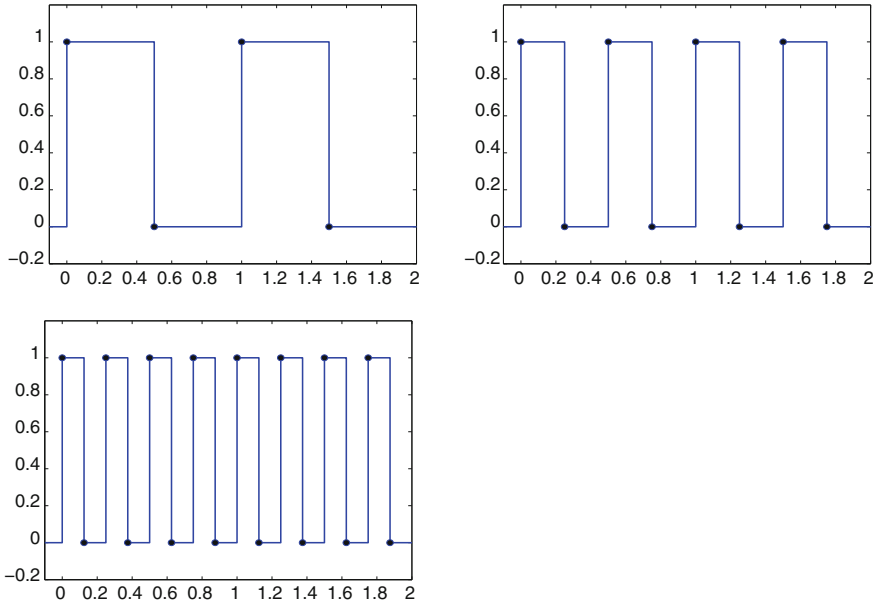
Fig. 10.2. $\phi_1 = Q_p\phi_0$, $\phi_2 = Q_p^2\phi_0$, $\phi_3 = Q_p^3\phi_0$ obtained by cascade algorithm with refinement mask $\{\frac{1}{2}, 0, \frac{1}{2}\}$ and $\phi_0 = \chi_{[0,1)}(x)$

Then, from ϕ_1, we obtain ϕ_2, given by

$$\phi_2(x) = Q_p\phi_1(x) = \phi_1(2x) + 0\phi_1(2x - 1) + \phi_1(2x - 2)$$
$$= \begin{cases} 1, & \text{for } x \in [0, 0.25) \cup [0.5, 0.75) \cup [1, 1.25) \cup [1.5, 1.75), \\ 0, & \text{elsewhere}; \end{cases}$$

and from ϕ_2, we obtain ϕ_3, ϕ_4, …. In this way, we can compute ϕ_n for $n \geq 3$. The graphs of ϕ_1, ϕ_2, ϕ_3 are displayed in Fig. 10.2. However, from the graphical results, it appears that ϕ_n does not converge to ϕ in $L_2(\mathbb{R})$. ∎

From the above two examples, we see that it is imperative to derive the mathematical theory to assure the convergence of the cascade algorithm. In Theorem 3, to be proved at the end of this section, we will see that the two sufficient conditions for the convergence of the cascade algorithm are: firstly, the finite refinement mask p has the sum-rule property at least of the first order; and secondly, the transition operator T_p associated with p satisfies Condition E. Furthermore, from the two examples that will be presented right after the statement of Theorem 3, we will see that neither of these conditions can be dropped to ensure convergence of the cascade algorithm. Before going into the discussion of the convergence of the cascade algorithm, we need some preliminary results. First let us demonstrate, in the following example, that for the cascade algorithm, with a compactly supported initial function ϕ_0, such that

supp(ϕ_0) $\subseteq$ [0, N], then for each $n = 1, 2, \ldots$, we also have supp(ϕ_n) $\subseteq$ [0, N], but the support generally increases for increasing values of n. Note that the refinable function ϕ of the refinement equation defined by p, being the limit of the sequence ϕ_n, will then have support, supp(ϕ) $\subseteq$ [0, N], as assured by Theorem 2 of Sect. 10.1.

Example 3 Let $p = \{p_k\}_{k=0}^{N}$ with non-zero p_0, p_N, and let $\phi_0 \in L_2(\mathbb{R})$, with supp($\phi_0$) $\subseteq$ [0, $N - \delta$], where $0 < \delta \leq 1$, be an initial function for the cascade algorithm, then

$$\text{supp}(\phi_n) \subseteq [0, \left(1 - (\tfrac{1}{2})^n\right)N] + (\tfrac{1}{2})^n \text{supp}(\phi_0). \tag{10.3.4}$$

This will be derived in the proof of Theorem 1 below. Hence, since supp(ϕ_0) $\subseteq$ [0, $N - \delta$], we have

$$\text{supp}(\phi_n) \subseteq [0, N - \frac{\delta}{2^n}].$$ ∎

Observe that if we choose an arbitrary compactly supported function $\phi_0 \in L_2(\mathbb{R})$ as an initial function for the cascade algorithm, then the support of ϕ_n may not be a subset of the interval [0, N], although the support of the limit function ϕ is (assuming that the cascade algorithm converges). For fast convergence, it is advisable to select an initial function ϕ_0 with a small support.

Next, recall that a compactly supported function $f(x)$ is said to have the property of partition of unity (PU) if it satisfies

$$\sum_{\ell \in \mathbb{Z}} f(x - \ell) = 1, \ x \in \mathbb{R}.$$

In the following Theorem 1, we will also prove that the PU is preserved by the cascade algorithm, provided that the refinement mask p has the sum-rule property of at least the first order, namely, $p(\pi) = 0$.

Theorem 1 $\boxed{\textbf{Support and PU of } \phi_n = Q_p^n \phi_0}$ *Let $\phi_n = Q_p^n \phi_0$, where ϕ_0 is any compactly supported function in $L_2(\mathbb{R})$. Then:*

(i) *There exists an $n_0 \geq 1$ such that supp(ϕ_n) $\subseteq [-\frac{1}{2}, N + \frac{1}{2}]$ for all $n \geq n_0$.*
(ii) *In addition, if p has at least sum-rule order 1; that is, $p(\pi) = 0$, and if the initial function $\phi_0(x)$ has PU, then each of $\phi_n(x), n = 1, 2, \ldots$, has PU.*

Proof To prove (i), observe that it follows from the definition of Q_p, given by (10.3.3), that

$$\text{supp}(\phi_1) = \text{supp}(Q_p \phi_0) \subseteq \bigcup_{k \in [0,N]} \text{supp}(\phi_0(2 \cdot -k))$$

$$= \bigcup_{k \in [0,N]} \frac{1}{2}(\text{supp}(\phi_0) + k) \subseteq \frac{1}{2}(\text{supp}(\phi_0) + [0, N])$$

$$= \frac{1}{2}[0, N] + \frac{1}{2}\text{supp}(\phi_0).$$

In general, we have

$$\text{supp}(\phi_n) \subseteq \frac{1}{2}[0, N] + \frac{1}{2^2}[0, N] + \cdots + \frac{1}{2^n}[0, N] + \frac{1}{2^n}\text{supp}(\phi_0).$$

Since ϕ_0 is compactly supported, an integer n_0 can be chosen sufficiently large so that $\frac{1}{2^n}\text{supp}(\phi_0) \subseteq [-\frac{1}{2}, \frac{1}{2}]$ for all $n \geq n_0$. Thus, we have

$$\text{supp}(\phi_n) \subseteq \sum_{j=1}^{\infty} \frac{1}{2^j}[0, N] + [-\frac{1}{2}, \frac{1}{2}] = [-\frac{1}{2}, N + \frac{1}{2}],$$

for $n \geq n_0$, as desired.

To prove (ii), we start by observing that $\widehat{\phi}_1(0) = p(0)\widehat{\phi}_0(0) = 1$. For an integer $k \neq 0$,

$$\widehat{\phi}_1(2\pi k) = p(\pi k)\widehat{\phi}_0(\pi k) = \begin{cases} p(0)\widehat{\phi}_0(2\pi\ell), & \text{for } k = 2\ell, \\ p(\pi)\widehat{\phi}_0(2\pi\ell + \pi), & \text{for } k = 2\ell + 1, \end{cases}$$

$$= 0,$$

since $p(\pi) = 0$ and $\widehat{\phi}_0(2\pi\ell) = 0$ for $\ell \neq 0$. Thus ϕ_1 has PU. By repeating this argument, we may deduce that $\phi_2, \phi_3, \ldots$, all have PU. ∎

From Theorem 5 in Sect. 10.1 on p.507, we also know that if a refinable function ϕ is in $L_2(\mathbb{R})$, then ϕ has PU. Next, we show that if the sequence $\{Q_p^n \phi_0\}_n$ converges in $L_2(\mathbb{R})$, then ϕ_0 must have PU.

Theorem 2 | **Convergence of $\{Q_p^n \phi_0\}_{n=1}^{\infty} \Rightarrow$ PU for initial ϕ_0** | *If, for a compactly supported initial function ϕ_0 with $\widehat{\phi}_0(0) = 1$, the cascade sequence $\phi_n = Q_p^n \phi_0$ converges in $L_2(\mathbb{R})$, then the initial function ϕ_0 must have the PU.*

Proof Since the sequence ϕ_n converges in $L_2(\mathbb{R})$, the sequence $\widehat{\phi}_n$ also converges in $L_2(\mathbb{R})$ by Parseval's theorem for the Fourier transform (see Theorem 1 on p.330). On the other hand, $\widehat{\phi}_n$ converges pointwise to $\widehat{\phi}$ on $\mathbb{R}$. Thus, $\widehat{\phi}_n$ converges in $L_2(\mathbb{R})$ to $\widehat{\phi}$. Hence, the sequence ϕ_n converges in $L_2(\mathbb{R})$ to ϕ. In addition, by (i) of Theorem 1, there exists an n_0 such that $\text{supp}(\phi_n) \subseteq [-\frac{1}{2}, N + \frac{1}{2}]$ for $n \geq n_0$. Thus, for $n \geq n_0$,

$$|\widehat{\phi}_n(\omega) - \widehat{\phi}(\omega)| \le \int_{-\infty}^{\infty} |\phi_n(x) - \phi(x)| dx = \int_{-\frac{1}{2}}^{N+\frac{1}{2}} |\phi_n(x) - \phi(x)| dx$$

$$\le (N+1)^{\frac{1}{2}} \|\phi_n - \phi\|_2 \to 0, \text{ as } n \to \infty,$$

for any $\omega \in \mathbb{R}$. This means that $\widehat{\phi}_n$ converges to $\widehat{\phi}$ uniformly on $\mathbb{R}$.

For any integer $k \ne 0$, observe that

$$\widehat{\phi}_n(2^n 2\pi k) = p(2^{n-1} 2\pi k) \widehat{\phi}_{n-1}(2^{n-1} 2\pi k) = p(0) \widehat{\phi}_{n-1}(2^{n-1} 2\pi k)$$
$$= \widehat{\phi}_{n-1}(2^{n-1} 2\pi k) = \cdots = \widehat{\phi}_0(2\pi k),$$

so that

$$\widehat{\phi}_0(2\pi k) = \lim_{n \to \infty} \widehat{\phi}_n(2^n 2\pi k) = \lim_{n \to \infty} \widehat{\phi}(2^n 2\pi k) = 0,$$

where the last equality holds by applying the Riemann-Lebesgue lemma. This shows ϕ_0 has the PU.   ∎

Based on Theorem 2, we introduce the following definition of the convergence of the cascade algorithm.

Definition 1 | **Convergence of cascade algorithm** | *The cascade algorithm associated with a refinement mask $p = \{p_k\}_{k=0}^{N}$ is said to be convergent in $L_2(\mathbb{R})$, if the sequence $\phi_n = \{Q_p^n \phi_0\}$ converges in $L_2(\mathbb{R})$, for any compactly supported initial function $\phi_0 \in L_2(\mathbb{R})$ that has the PU.*

The main theorem of this section is the following.

Theorem 3 | **Characterization of convergence of cascade algorithm** | *The cascade algorithm associated with $p = \{p_k\}_{k=0}^{N}$ converges in $L_2(\mathbb{R})$ to ϕ if and only if*

 (i) *p has at least sum-rule order 1; and*
(ii) *$T_p \in E$.*

The proof of Theorem 3 will be given at the end of this section. Let us first point out that, by Theorem 3, we see that the cascade algorithm associated with the refinement mask $p = \{p_k\}_{k=0}^{2}$ of the hat function, as considered in Example 1, converges in $L_2(\mathbb{R})$ to the hat function, since p has sum-rule order 2 and T_p satisfies Condition E (see Example 2 on p.443).

Example 4 Let $p = \{p_k\}_{k=0}^{2}$ with $p_0 = p_2 = 1$, $p_1 = 0$ be the refinement mask of $\phi(x) = \frac{1}{2}\chi_{[0,2)}(x)$ considered in Example 2. Discuss the convergence of the associated cascade algorithm.

Solution The two-scale symbol is given by

$$p(\omega) = \frac{1}{2}(1 + e^{-i2\omega}).$$

Since $p(\omega)$ does not have sum-rule property of any order, we immediately know, from Theorem 3, that the associated cascade algorithm could not converge in $L_2(\mathbb{R})$. In the following, let us consider this problem in more depth.

Firstly, let us consider whether $T_p \in E$. From Example 1 of Sect. 10.1, on p.505, we have

$$|\lambda I_5 - T_p| = (\lambda - \frac{1}{2})^2 \begin{vmatrix} \lambda & -\frac{1}{2} & 0 \\ 0 & \lambda - 1 & 0 \\ 0 & -\frac{1}{2} & \lambda \end{vmatrix} = (\lambda - 1)(\lambda - \frac{1}{2})^2 \lambda^2.$$

Thus, the eigenvalues of T_p are given by

$$1, \frac{1}{2}, \frac{1}{2}, 0, 0;$$

from which we may conclude that T_p satisfies Condition E.

In the following, we show directly that the cascade algorithm does not converge. Observe that if ϕ_n is supported on $[0, 2]$, then

$$\text{supp}(\phi_n(2\cdot)) \subseteq [0, 1], \ \text{supp}(\phi_n(2\cdot-2)) \subseteq [1, 2];$$

and ϕ_{n+1}, obtained by the cascade algorithm, given by

$$\phi_{n+1}(x) = \phi_n(2x) + \phi_n(2x - 2),$$

is also supported on $[0, 2]$. Thus, for ϕ_0 supported on $[0, 2]$, $\text{supp}(\phi_n) \subseteq [0, 2]$ for $n \geq 1$, and $\text{supp}(\phi_n(2\cdot)) \cap \text{supp}(\phi_n(2\cdot-2)) \subseteq \{1\}$. Hence we have

$$\int_{-\infty}^{\infty} |\phi_{n+1}(x)|^2 dx = \int_{-\infty}^{1} |\phi_{n+1}(x)|^2 dx + \int_{1}^{\infty} |\phi_{n+1}(x)|^2 dx$$

$$= \int_{-\infty}^{\infty} |\phi_n(2x)|^2 dx + \int_{-\infty}^{\infty} |\phi_n(2x - 2)|^2 dx$$

$$= \frac{1}{2} \int_{-\infty}^{\infty} |\phi_n(x)|^2 dx + \frac{1}{2} \int_{-\infty}^{\infty} |\phi_n(x)|^2 dx = \int_{-\infty}^{\infty} |\phi_n(x)|^2 dx$$

$$= \cdots = \int_{-\infty}^{\infty} |\phi_0(x)|^2 dx.$$

In particular, by selecting $\phi_0(x) = \chi_{[0,1)}(x)$ as the initial function, we have

$$\lim_{n \to \infty} \|\phi_n\|_2 = \lim_{n \to \infty} \|\phi_0\|_2 = 1. \tag{10.3.5}$$

On the other hand, if the cascade algorithm associated with p converges in $L_2(\mathbb{R})$, then

$$\|\phi_n\|_2 \to \|\phi\|_2 = \frac{\sqrt{2}}{2} \text{ as } n \to \infty,$$

and this contradicts (10.3.5), showing that ϕ_n does not converge in $L_2(\mathbb{R})$ to ϕ.   ∎

The above example demonstrates that the sum-rule property is a necessary condition for the validity of Theorem 3. In the next example, we show that Condition E is also necessary.

Example 5 Let $\phi(x) = \frac{1}{3}\chi_{[0,3)}(x)$ be the refinable function considered in Example 4 on p.450 and in Example 2 on p.508. Discuss the convergence of the associated cascade algorithm.

Solution The two-scale symbol is given by

$$p(\omega) = \frac{1}{2}(1 + e^{-i3\omega}).$$

The entries a_j of $\mathcal{T}_p$ can be obtained from

$$\sum_j a_j e^{-ij\omega} = 4|p(\omega)|^2 = 2 + e^{-i3\omega} + e^{i3\omega},$$

with the nonzero values given by

$$a_0 = 2, a_3 = a_{-3} = 1,$$

and

$$\mathcal{T}_p = \left[\frac{1}{2} a_{2j-k}\right]_{-3 \leq j,k \leq 3} = \frac{1}{2} \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 2 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

Also, it can be shown that

$$|\lambda I_7 - \mathcal{T}_p| = (\lambda - 1)^2 (\lambda + 1)(\lambda - \frac{1}{2})^3 (\lambda + \frac{1}{2}).$$

Thus, the eigenvalues of $\mathcal{T}_p$ are given by

$$1, 1, -1, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, -\frac{1}{2}.$$

Hence, T_p does not satisfy Condition E, and Theorem 3 does not apply to conclude the convergence of the cascade algorithm. As in Example 4, let us demonstrate the divergence of the cascade algorithm directly.

First, notice that if $\text{supp}(\phi_n) \subseteq [0, 3]$, then

$$\text{supp}(\phi_n(2\cdot)) \subseteq [0, \frac{3}{2}], \ \text{supp}(\phi_n(2 \cdot -3)) \subseteq [\frac{3}{2}, 3],$$

and hence ϕ_{n+1}, obtained by the cascade algorithm, given by

$$\phi_{n+1}(x) = \phi_n(2x) + \phi_n(2x - 3),$$

is also supported on $[0, 3]$. Thus, for ϕ_0 supported on $[0, 3]$, we have

$$\int_{-\infty}^{\infty} |\phi_{n+1}(x)|^2 dx = \int_{-\infty}^{\infty} |\phi_n(2x)|^2 dx + \int_{-\infty}^{\infty} |\phi_n(2x - 3)|^2 dx$$

$$\text{(by the fact } \text{supp}(\phi_n(2\cdot)) \cap \text{supp}(\phi_n(2 \cdot -3)) \subseteq \{\frac{3}{2}\})$$

$$= \frac{1}{2} \int_{-\infty}^{\infty} |\phi_n(x)|^2 dx + \frac{1}{2} \int_{-\infty}^{\infty} |\phi_n(x)|^2 dx = \int_{-\infty}^{\infty} |\phi_n(x)|^2 dx$$

$$= \cdots = \int_{-\infty}^{\infty} |\phi_0(x)|^2 dx.$$

If the cascade algorithm is convergent, then

$$\|\phi_n\|_2 \to \|\phi\|_2, \text{ as } n \to \infty.$$

However, for $\phi_0(x) = \chi_{[0,1)}(x)$ or $\phi_0(x) = \max\{x, 2 - x\}\chi_{[0,2)}$ (the hat function, where the verification of its PU is left as an exercise),

$$\lim_{n\to\infty} \|\phi_n\|_2 = \lim_{n\to\infty} \|\phi_0\|_2 \neq \frac{\sqrt{3}}{3} = \|\phi\|_2,$$

a contradiction. This shows that ϕ_n does not converge in $L_2(\mathbb{R})$ to ϕ. In Fig. 10.3, we present ϕ_1 and ϕ_4, with ϕ_0 being the hat function. ∎

From the above two examples, we see both conditions (i) and (ii) in Theorem 3 are necessary for the convergence of the cascade algorithm.

By Theorem 1, if a refinable function ϕ is stable, its associated mask p must have at least sum-rule order 1, and $T_p \in E$. Thus Theorem 3 leads to the following corollary.

Corollary 1 $\boxed{\textbf{Stability} \Rightarrow \textbf{convergence of cascade algorithm}}$ *If a refinable function ϕ is stable (i.e., ϕ is a scaling function), then the cascade algorithm associated with its refinement mask converges.*
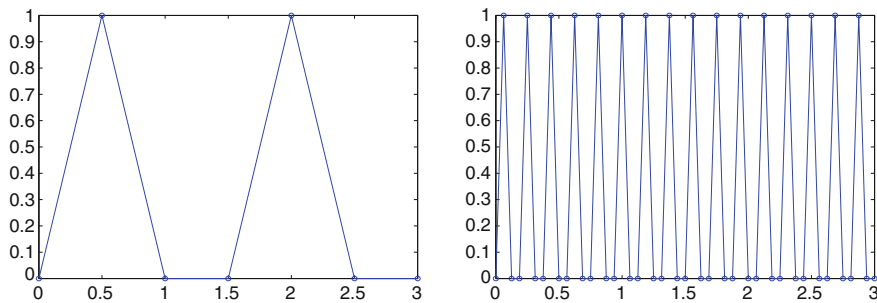
Fig. 10.3. $\phi_1 = Q_p \phi_0$, $\phi_4 = Q_p^4 \phi_0$ obtained by cascade algorithm with refinement mask of $\phi = \frac{1}{3}\chi_{[0,3)}$ and ϕ_0 being the hat function

Example 6   Let p be the D_4 filter with sum-rule order 2. Since the refinable function ϕ is orthogonal, it is stable. Thus, by Corollary 1, the cascade algorithm associated with p is convergent. Indeed, from Example 2 on p.509, T_p satisfies Condition E, and thus, we have the same conclusion on the convergence of the associated cascade algorithm. In particular, if we choose $\phi_0(x) = \min\{x, 2 - x\}\chi_{[0,2)}(x)$ (the hat function) as the initial function, then ϕ_n, $n = 1, 2, \ldots$, provide a sequence of piecewise linear polynomials that approximate ϕ. The graphs of ϕ_n with $n = 1, \ldots, 4$ are shown in Fig. 10.4. Efficient schemes, based on the cascade algorithm, can be used to draw (i.e. to render) the graphs of ϕ and the associated wavelet ψ, as well as graphs of functions written as finite linear combinations of their integer translates.

∎

Next, we apply the characterization of the convergence of the cascade algorithm to prove Theorem 5 on p.471 on the biorthogonality of two scaling functions.

Proof of Theorem 5 of Sect. 9.4 on p.471   Suppose that ϕ and $\widetilde{\phi}$ are biorthogonal to each other. Then item (i) in Theorem 5 has been proved in Theorem 4 on p.470, and items (ii) and (iii) follow from Theorem 1 of Sect. 10.2 on p.510, since the biorthogonality of ϕ and $\widetilde{\phi}$ immediately implies the stability of ϕ and $\widetilde{\phi}$, respectively (see Corollary 1 on p.467).

Conversely, suppose (i)–(iii) in Theorem 5 hold. By Theorem 3, the cascade algorithms associated with p and with $\widetilde{p}$ converge in $L_2(\mathbb{R})$. In particular, with $\phi_0(x) = \widetilde{\phi}_0(x) = \chi_{[0,1)}(x)$ as initial functions for both ϕ_n and $\widetilde{\phi}_n$; that is, by considering

$$\phi_n = (Q_p)^n \phi_0, \ \widetilde{\phi}_n = (Q_{\widetilde{p}})^n \widetilde{\phi}_0, \ n = 1, 2, \ldots,$$

ϕ_n converges to ϕ in $L_2(\mathbb{R})$, and $\widetilde{\phi}_n$ converges to $\widetilde{\phi}$ in $L_2(\mathbb{R})$. Thus, the refinable functions ϕ and $\widetilde{\phi}$ are in $L_2(\mathbb{R})$.

To show that ϕ and $\widetilde{\phi}$ are biorthogonal, we next show $G_{\widetilde{\phi}_n, \phi_n}(\omega) = 1, \omega \in [-\pi, \pi]$ for $n = 0, 1, \ldots$, by mathematical induction. Indeed, since $G_{\widetilde{\phi}_0, \phi_0}(\omega) = G_{\phi_0}(\omega) = 1$ (in view of the orthogonality of ϕ_0), we may use the induction hypothesis, $G_{\widetilde{\phi}_n, \phi_n}(\omega) = 1$. Now, by applying this assumption and the refinement relations
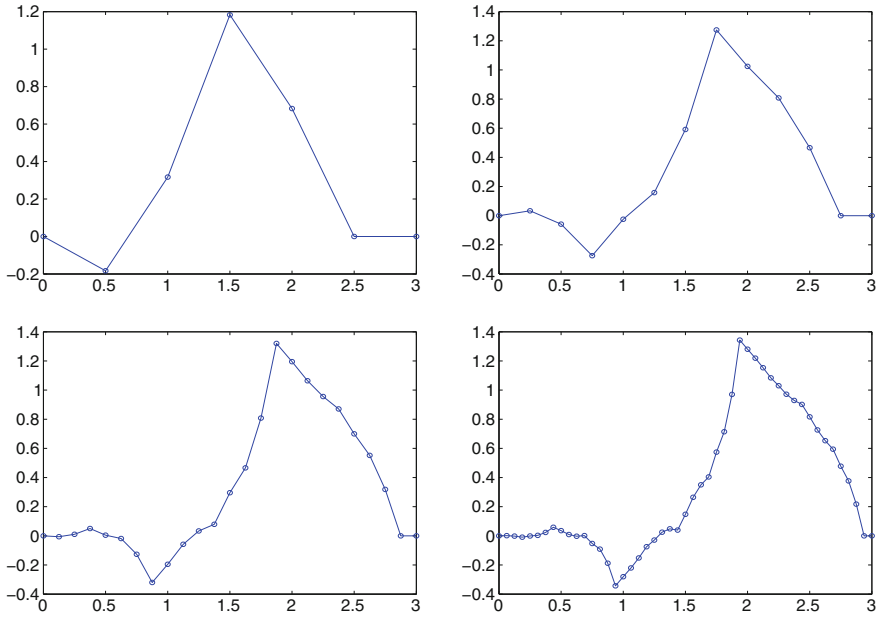
Fig. 10.4. Piecewise linear polynomials $\phi_n = Q_p^n \phi_0, n = 1, \ldots, 4$ approximating D_4 scaling function ϕ by cascade algorithm

in the calculation below, we arrive at

$$
\begin{aligned}
G_{\widetilde{\phi}_{n+1}, \phi_{n+1}}(\omega) &= \sum_{\ell \in \mathbb{Z}} \widehat{\widetilde{\phi}}_{n+1}(\omega + 2\pi\ell) \overline{\widehat{\phi}_{n+1}(\omega + 2\pi\ell)} \\
&= \sum_{\ell \in \mathbb{Z}} \widetilde{p}(\frac{\omega}{2} + \pi\ell) \widehat{\widetilde{\phi}}_n(\frac{\omega}{2} + \pi\ell) \overline{p(\frac{\omega}{2} + \pi\ell)} \overline{\widehat{\phi}_n(\frac{\omega}{2} + \pi\ell)} \\
&= \sum_{k \in \mathbb{Z}} \widetilde{p}(\frac{\omega}{2} + 2\pi k) \overline{p(\frac{\omega}{2} + 2\pi k)} \widehat{\widetilde{\phi}}_n(\frac{\omega}{2} + 2\pi k) \overline{\widehat{\phi}_n(\frac{\omega}{2} + 2\pi k)} \\
&\quad + \sum_{k \in \mathbb{Z}} \widetilde{p}(\frac{\omega}{2} + 2\pi k + \pi) \overline{p(\frac{\omega}{2} + 2\pi k + \pi)} \widehat{\widetilde{\phi}}_n(\frac{\omega}{2} + 2\pi k + \pi) \overline{\widehat{\phi}_n(\frac{\omega}{2} + 2\pi k + \pi)} \\
&= \widetilde{p}(\frac{\omega}{2}) \overline{p(\frac{\omega}{2})} G_{\widetilde{\phi}_n, \phi_n}(\frac{\omega}{2}) + \widetilde{p}(\frac{\omega}{2} + \pi) \overline{p(\frac{\omega}{2} + \pi)} G_{\widetilde{\phi}_n, \phi_n}(\frac{\omega}{2} + \pi) \\
&= \widetilde{p}(\frac{\omega}{2}) \overline{p(\frac{\omega}{2})} + \widetilde{p}(\frac{\omega}{2} + \pi) \overline{p(\frac{\omega}{2} + \pi)} = 1.
\end{aligned}
$$

This shows that $G_{\widetilde{\phi}_n, \phi_n}(\omega) = 1, \omega \in [-\pi, \pi]$ for all $n \geq 0$; that is, ϕ_n and $\widetilde{\phi}_n$ are biorthogonal to each other, namely:

$$
\int_{-\infty}^{\infty} \widetilde{\phi}_n(x) \overline{\phi_n(x - k)} dx = \delta_k, \ k \in \mathbb{Z}.
$$

Since ϕ_n and $\widetilde{\phi}_n$ converge in $L_2(\mathbb{R})$ to ϕ and $\widetilde{\phi}$, respectively, the limit of the left-hand side of the above equation, as $n \to \infty$, is $\int_{-\infty}^{\infty} \widetilde{\phi}(x)\overline{\phi(x-k)}dx$ (see Exercise 8). Thus we have

$$\int_{-\infty}^{\infty} \widetilde{\phi}(x)\overline{\phi(x-k)}dx = \delta_k, \; k \in \mathbb{Z}.$$

This shows that ϕ and $\widetilde{\phi}$ are biorthogonal to each other, as desired. ∎

Example 7 Let p and $\widetilde{p}$ be the 5/3-tap biorthogonal lowpass filters considered in Example 1 on p.475, with two-scale symbols:

$$p(\omega) = \frac{1}{4}(1 + e^{-i\omega})^2,$$

$$\widetilde{p}(\omega) = -\frac{1}{8}e^{i\omega} + \frac{1}{4} + \frac{3}{4}e^{-i\omega} + \frac{1}{4}e^{-i2\omega} - \frac{1}{8}e^{-i3\omega}.$$

Then it is not difficult to calculate the nonzero $a_j, \widetilde{a}_j$, which are

$$a_0 = \frac{3}{2}, \; a_1 = a_{-1} = 1, \; a_2 = a_{-2} = \frac{1}{4},$$

and

$$\widetilde{a}_0 = \frac{23}{8}, \; \widetilde{a}_1 = \widetilde{a}_{-1} = \frac{5}{4}, \; \widetilde{a}_2 = \widetilde{a}_{-2} = -\frac{1}{2},$$

$$\widetilde{a}_3 = \widetilde{a}_{-3} = -\frac{1}{4}, \; \widetilde{a}_4 = \widetilde{a}_{-4} = \frac{1}{16}.$$

Then we can compute the eigenvalues of $\mathcal{T}_p = \left[\frac{1}{2}a_{2j-k}\right]_{-2 \leq j,k \leq 2}$, which are given by

$$1, \; \frac{1}{2}, \; \frac{1}{4}, \; \frac{1}{8}, \; \frac{1}{8}.$$

The eigenvalues of $\mathcal{T}_{\widetilde{p}} = \left[\frac{1}{2}\widetilde{a}_{2j-k}\right]_{-4 \leq j,k \leq 4}$ can be calculated by using any computational software. For example, numerical results generated from Matlab are:

$$1, \; \frac{1}{2}, \; \frac{1}{4}, \; \frac{1}{8}, \; 0.54279, \; -\frac{1}{4}, \; -0.23029, \; \frac{1}{32}, \; \frac{1}{32}.$$

Thus, both T_p and $T_{\widetilde{p}}$ satisfy Condition E. In addition, p and $\widetilde{p}$ have at least sum-rule order 1. Therefore, by Theorem 5 on p.471, we may conclude that the associated refinable functions ϕ and $\widetilde{\phi}$ are biorthogonal to each other. ∎

Example 8 Let p and $\widetilde{p}$ be the biorthogonal lowpass filters given in (9.4.26) in Example 2 of Sect. 9.2 on p.447. The largest eigenvalue of

$$T_{\widetilde{p}} = \left[\frac{1}{2}\,\widetilde{a}_{2j-k}\right]_{-4\leq j,k\leq 4}$$

is 3.1378. Thus $T_{\widetilde{p}}$ does not satisfy Condition E, and hence the associated refinable functions ϕ and $\widetilde{\phi}$ are not biorthogonal to each other. ∎

We end this section by giving the proof of the main theorem of the section, Theorem 3.

Proof of Theorem 3 Suppose the cascade algorithm associated with p converges in $L_2(\mathbb{R})$. To prove item (i) in Theorem 3, since we only consider finite sequences that sum to 2, we already have $p(0) = 1$; that is, it is sufficient to show $p(\pi) = 0$. Choose the initial function $\phi_0(x) = \chi_{[0,1)}(x)$. Then ϕ_0 has PU, and hence, the sequence $\{\phi_n\}$, with $\phi_n = Q_p^n \phi_0$, converges in $L_2(\mathbb{R})$ to ϕ. On the other hand, as shown in the proof of Theorem 2, we know that $\{\widehat{\phi}_n\}$ converges to $\widehat{\phi}$ uniformly on $\mathbb{R}$. Thus, with

$$\widehat{\phi}_n(2^n\pi) = \widehat{\phi}_{n-1}(2^{n-1}\pi) = \cdots = \widehat{\phi}_1(2\pi) = p(\pi)\widehat{\phi}_0(\pi),$$

we have

$$p(\pi)\widehat{\phi}_0(\pi) = \lim_{n\to\infty}\widehat{\phi}_n(2^n\pi) = \lim_{n\to\infty}\widehat{\phi}(2^n\pi) = 0,$$

where the last equality follows by applying the Riemann-Lebesgue lemma. This, together with the fact that $\widehat{\phi}_0(\pi) = \frac{1-e^{-i\pi}}{i\pi} = \frac{2}{i\pi} \neq 0$, leads to $p(\pi) = 0$.

To prove item (ii) in Theorem 3, we note that $G_{\phi_n}(\omega) = T_p^n G_{\phi_0}(\omega)$ (see Exercise 5). Thus, by (9.2.6) on p.447, for a compactly supported $\phi_0 \in L_2(\mathbb{R})$ with PU, we have

$$
\begin{aligned}
T_p^n G_{\phi_0}(\omega) = G_{\phi_n}(\omega) &= \sum_{k=-(N-1)}^{N-1} \int_{-\infty}^{\infty} \phi_n(x)\overline{\phi_n(x-k)}dx\, e^{-ik\omega} \\
&\to \sum_{k=-(N-1)}^{N-1} \int_{-\infty}^{\infty} \phi(x)\overline{\phi(x-k)}dx\, e^{-ik\omega} \text{ as} n \to \infty \\
&= G_\phi(\omega),
\end{aligned}
$$

for $\omega \in [-\pi, \pi]$, where

$$\lim_{n\to\infty}\int_{-\infty}^{\infty}\phi_n(x)\overline{\phi_n(x-k)}dx = \int_{-\infty}^{\infty}\phi(x)\overline{\phi(x-k)}dx$$

follows from the fact that $\|\phi_n - \phi\|_2 \to 0$ as $n \to \infty$ (see Exercise 8). Thus,

$$\lim_{n\to\infty} T_p^n G_{\phi_0}(\omega) = G_\phi(\omega), \ \omega \in [-\pi, \pi]. \tag{10.3.6}$$

For each j, with $-N \leq j \leq N$, let us choose the initial function $\phi_0(x) = \chi_{[j, j+1)}(x)$. Then ϕ_0 has PU. Thus, with $G_{\phi_0}(\omega) = e^{-ij\omega}$ (see Exercise 6), and from (10.3.6), we have

$$\lim_{n \to \infty} T_p^n(e^{-ij\omega}) = G_\phi(\omega).$$

Therefore, for any $v(\omega) = \sum_{k=-N}^{N} v_k e^{-ik\omega} \in \mathbb{V}_{2N+1}$,

$$\lim_{n \to \infty} T_p^n v(\omega) = \Big(\sum_{k=-N}^{N} v_k \Big) G_\phi(\omega), \ \omega \in [-\pi, \pi]. \tag{10.3.7}$$

Since, for $v(\omega) \in \mathbb{V}_{2N+1}$, $T_p^n v(\omega)$ converges for any $\omega \in [-\pi, \pi]$, the eigenvalue λ of T_p must satisfy either $|\lambda| < 1$ or $\lambda = 1$, and the eigenvalue 1 is nondegenerate.

Next, we show that 1 is a simple eigenvalue. Let $v(\omega) \in \mathbb{V}_{2N+1}$ be a 1-eigenfunction of T_p. Then, by (10.3.7), we have

$$v(\omega) = \lim_{n \to \infty} 1^n \, v(\omega) = \lim_{n \to \infty} T_p^n v(\omega) = \Big(\sum_{k=-N}^{N} v_k \Big) G_\phi(\omega).$$

Thus, up to a constant multiple, $G_\phi(\omega)$ is the only 1-eigenfunction of T_p. This shows $T_p \in \mathrm{E}$.

Conversely, suppose that p has at least sum-rule order 1 and that $T_p \in \mathrm{E}$. For a compactly supported ϕ_0 with PU, we know, by (i) in Theorem 1, that there is a positive integer n_0 such that $\mathrm{supp}(\phi_n) \subseteq [-\frac{1}{2}, N + \frac{1}{2}]$ for $n \geq n_0$. In addition, by (ii) in Theorem 1, it follows that ϕ_{n_0} has PU. By considering the convergence of $\phi_{n+n_0} = Q_p^n(\phi_{n_0})$ with ϕ_{n_0} as the initial function, we may assume that the support of ϕ_0 is in $[-\frac{1}{2}, N + \frac{1}{2}]$. Then, for $n \geq 1$, $\mathrm{supp}(\phi_n) \subseteq [-\frac{1}{2}, N + \frac{1}{2}]$ (by referring to the proof of (i) in Theorem 1), and hence, $G_{\phi_n} \in \mathbb{V}_{2N+1}$.

From the fact that $G_{\phi_n}(\omega) = T_p^n G_{\phi_0}(\omega)$ and the assumption that $T_p \in \mathrm{E}$, we know that $G_{\phi_n}(\omega)$ converges pointwise on $[-\pi, \pi]$. Let $W(\omega)$ denote the limit. By (ii) in Theorem 1, ϕ_n has PU. Thus,

$$G_{\phi_n}(0) = \sum_{\ell} |\widehat{\phi}_n(2\pi\ell)|^2 = 1.$$

Therefore, $W(0) = 1$. Clearly, $W(\omega) \geq 0$. Thus, by Theorem 3 on p.504, the refinable function ϕ is in $L_2(\mathbb{R})$. Therefore $G_\phi(\omega) \in \mathbb{V}_{2N+1}$, and it is a 1-eigenfunction of T_p. Since $G_{\phi_{n+1}}(\omega) = T_p G_{\phi_n}(\omega)$, we have

$$T_p W(\omega) = W(\omega);$$

that is, $W(\omega)$ is also a 1-eigenfunction of T_p. Since 1 is a simple eigenvalue of T_p, there exists a constant c such that

$$W(\omega) = c\, G_\phi(\omega).$$

In view of $W(0) = 1$ and $G_\phi(0) = 1$, we may deduce $c = 1$. Hence, we have $W(\omega) = G_\phi(\omega)$. This means that $G_{\phi_n}(\omega)$ converges pointwise to $G_\phi(\omega)$ on $[-\pi, \pi]$; that is,

$$\sum_{k=-N}^{N} \int_{-\infty}^{\infty} \phi_n(x)\overline{\phi_n(x-k)}dx\, e^{-ik\omega} \;\to\; \sum_{k=-N}^{N} \int_{-\infty}^{\infty} \phi(x)\overline{\phi(x-k)}dx\, e^{-ik\omega}$$

as $n \to \infty$ for $\omega \in [-\pi, \pi]$. In particular, we have

$$\int_{-\infty}^{\infty} |\phi_n(x)|^2 dx \;\to\; \int_{-\infty}^{\infty} |\phi(x)|^2 dx \text{ as } n \to \infty.$$

Therefore, by Parseval's formula (7.2.7) in Theorem 2 on p.331, we have

$$\|\widehat{\phi}_n\|_2 \to \|\widehat{\phi}\|_2 \text{ as } n \to \infty.$$

This, together with the fact that $\widehat{\phi}_n(\omega)$ converges pointwise on $\mathbb{R}$ to $\widehat{\phi}(\omega)$, implies that $\widehat{\phi}_n(\omega)$ converges to $\widehat{\phi}(\omega)$ in $L_2(\mathbb{R})$ (by applying a standard result in Real Analysis). Hence, by applying Parseval's formula again, we may conclude that $\phi_n(x)$ converges to $\phi(x)$ in $L_2(\mathbb{R})$. ∎

Exercises

Exercise 1 Let $p = \{p_k\}_{k=0}^{2}$ with

$$p_0 = \frac{1}{2},\ p_1 = 1,\ p_2 = \frac{1}{2}.$$

Show that the vector $\mathbf{y}_0 = [0, 1, 0]^T$ is a 1-eigenvector of the matrix $\mathcal{P}_3$ in (10.3.1). Then apply the algorithm described by (10.3.2) to compute the values of the refinement function $\phi(x)$ with refinement mask p, for $x = 0, 1/16, 2/16, \ldots, 31/16, 2$. Observe that $\phi(x)$ is the hat function, $\phi(x) = \min\{x, 2 - x\}\chi_{[0,2)}(x)$, called the linear (Cardinal) B-spline.

Exercise 2 Let $p = \{p_k\}_{k=0}^{3}$ with

$$p_0 = \frac{1}{4},\ p_1 = p_2 = \frac{3}{4},\ p_3 = \frac{1}{4}.$$

Show that the vector $\mathbf{y}_0 = [0, \frac{1}{2}, \frac{1}{2}, 0]^T$ is a 1-eigenvector of the matrix $\mathcal{P}_4$ in (10.3.1). Then apply the algorithm described by (10.3.2) to compute the values of the refinement function $\phi(x)$ with refinement mask p, for $x = 0, 1/16, 2/16, \ldots, 47/16, 3$. We remark that $\phi(x)$ is called the quadratic (Cardinal) B-spline.

Exercise 3 Let $p = \{p_k\}_{k=0}^4$ with

$$p_0 = \frac{1}{8}, \; p_1 = \frac{4}{8}, \; p_2 = \frac{6}{8}, \; p_3 = \frac{4}{8}, \; p_4 = \frac{1}{8}.$$

Show that the vector $\mathbf{y}_0 = [0, \frac{1}{6}, \frac{4}{6}, \frac{1}{6}, 0]^T$ is a 1-eigenvector of the matrix $\mathcal{P}_5$ in (10.3.1). Then apply the algorithm described by (10.3.2) to compute the values of the refinement function $\phi(x)$ with refinement mask p, for $x = 0, 1/16, 2/16, \ldots,$ $63/16, 4$. We remark that $\phi(x)$ is called the cubic (Cardinal) B-spline.

Exercise 4 Let $p = \{p_k\}_{k=0}^2$ with

$$p_0 = \frac{3}{5}, \; p_1 = 1, \; p_2 = \frac{2}{5},$$

be a refinement mask. Find $\phi_1 = Q_p \phi_0$, $\phi_2 = Q_p^2 \phi_0$, where $\phi_0(x) = \chi_{[0,1)}(x)$.

Exercise 5 Let $\phi_n = Q_p^n \phi_0$, $n = 1, 2, \ldots$, where Q_p is the refinement opera-tor defined by (10.3.3). Prove that $G_{\phi_1}(\omega) = T_p G_{\phi_0}(\omega)$, and then conclude that $G_{\phi_n}(\omega) = T_p^n G_{\phi_0}(\omega)$.

Exercise 6 Let $\phi_0(x) = \chi_{[j, j+1)}(x)$. Show that $G_{\phi_0}(\omega) = e^{-ij\omega}$.

Exercise 7 Let $\phi_0(x) = \min\{x, 2 - x\}\chi_{[0,2)}(x)$ be the hat function (also called the linear Cardinal B-spline). Show that ϕ_0 has the property of partition of unity (PU).

Exercise 8 Show that if $\| f_n - f \|_2 \to 0$ and $\|g_n - g\|_2 \to 0$ as $n \to \infty$, then

$$\lim_{n \to \infty} \int_{-\infty}^{\infty} f_n(x)\overline{g_n(x)}dx = \int_{-\infty}^{\infty} f(x)\overline{g(x)}dx.$$

Exercise 9 Let $\mathcal{T}_p$ be the matrix in Example 5. Verify that its eigenvalues are

$$1, 1, -1, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, -\frac{1}{2}.$$

Exercise 10 Let $p = \{p_k\}_{k=0}^2$ with

$$p_0 = \frac{3}{4}, \; p_1 = 1, \; p_2 = \frac{1}{4},$$

be the refinement mask considered in Example 1 on p.511. Decide whether the cascade algorithm associated with p is convergent in $L_2(\mathbb{R})$.

Exercise 11 Repeat Exercise 10 for $p = \{p_k\}_{k=0}^2$ with

$$p_0 = \frac{1}{3}, \; p_1 = 1, \; p_2 = \frac{2}{3}.$$

10.4 Smoothness of Compactly Supported Wavelets

In the previous section, we have studied how the graphs of refinable functions and any (finite) linear combination of their integer translates, including the corresponding wavelets, can be drawn, but there was no indication of the quality of these graphs. The objective of this section is to characterize the smoothness of compactly supported refinable functions ϕ (and hence their corresponding wavelets) in terms of the refinement masks (or sequences). In this book, we only study the quality of smoothness measured by Sobolev regularity, and will calibrate the Sobolev regularity (or smoothness) of a refinement function ϕ in terms of the eigenvalues of the transition operator T_p, associated with the corresponding refinement mask p of ϕ.

For any $\gamma \geq 0$, a function f is said to be in the Sobolev space $\mathbb{W}^{\gamma}(\mathbb{R})$ if the Fourier transform of f satisfies the condition

$$(1 + \omega^2)^{\frac{\gamma}{2}} \widehat{f}(\omega) \in L_2(\mathbb{R}).$$

In the following, the supremum of all values of γ, denoted by

$$\nu_f = \sup\{\gamma : f \in \mathbb{W}^{\gamma}(\mathbb{R})\},$$

is called the **critical Sobolev exponent**.

The issue that must be addressed is the reason for studying the Sobolev regularity instead of the more natural Hölder regularity of the same "order" $\gamma > 0$, defined by the space of functions of $f \in \mathbb{C}^k(\mathbb{R})$ with

$$|f^{(k)}(x) - f^{(k)}(y)| \leq C|x - y|^{\alpha}, \text{ for all } x, y \in \mathbb{R},$$

where $0 < \alpha = \gamma - k < 1$, for some positive constant C. While the derivation of the Hölder regularity is not more difficult than that of the Sobolev regularity to be discussed in this section, we prefer a unified theory based on the transition operator T_p associated with the refinement mask p, since the Sobolev regularity estimate provided in this section is characterized by the eigenvalues of T_p. The advantage of the Sobolev regularity over the Hölder regularity is its computational efficiency. This will be studied in this section with four demonstrative examples. Observe that if $f^{(k)}(x) \in L_2(\mathbb{R})$, then $\omega^k \widehat{f}(\omega) \in L_2(\mathbb{R})$ by applying Plancherel's formula (called Parseval's theorem in Theorem 1 of Sect. 7.2), namely:

$$\|\omega^k \widehat{f}(\omega)\|_2^2 = 2\pi \|f^{(k)}\|_2^2.$$

Thus, if in addition, $f(x) \in L_2(\mathbb{R})$, then $f \in \mathbb{W}^k(\mathbb{R})$. On the other hand, it can also be proved that if $f \in \mathbb{W}^\gamma(\mathbb{R})$ for an arbitrary $\gamma > k + \frac{1}{2}$, where k is a given nonnegative integer, then $f \in \mathbb{C}^k(\mathbb{R})$. Thus, the Sobolev regularity assures smoothness of functions in a similar way as the Hölder regularity.

Let $\phi \in L_2(\mathbb{R})$ be the refinable function associated with a refinement mask $p = \{p_k\}_{k=0}^N$, which has sum-rule order $L \geq 1$. Then the refinement mask $p(\omega)$ is a trigonometric polynomial that can be written as

$$p(\omega) = \Big(\frac{1 + e^{-i\omega}}{2}\Big)^L p_0(\omega) = \Big(\cos\frac{\omega}{2}\Big)^L e^{-i\frac{\omega}{2}L} p_0(\omega),$$

for some trigonometric polynomial $p_0(\omega)$. Let $\mathbb{U}_{2L}$ denote the subspace of $\mathbb{V}_{2N+1}$ defined by

$$\mathbb{U}_{2L} = \Big\{u(\omega) \in \mathbb{V}_{2N+1} : \frac{d^\ell}{d\omega^\ell}u(0) = 0, \text{ for all } 0 \leq \ell \leq 2L - 1\Big\}. \quad (10.4.1)$$

Theorem 1 *If p has sum-rule order L, then the subspace $\mathbb{U}_{2L}$ is invariant under T_p.*

Proof　Any $u(\omega) \in \mathbb{U}_{2L}$ can be written as

$$u(\omega) = \Big(\frac{1 - e^{-i\omega}}{2i}\Big)^{2L} u_0(\omega) = \Big(\sin\frac{\omega}{2}\Big)^{2L} e^{-i\omega L} u_0(\omega),$$

where $u_0(\omega)$ is a trigonometric polynomial. Thus

$$\begin{aligned}
T_p u(\omega) &= |p(\frac{\omega}{2})|^2 u(\frac{\omega}{2}) + |p(\frac{\omega}{2} + \pi)|^2 u(\frac{\omega}{2} + \pi)\\
&= |p(\frac{\omega}{2})|^2 \Big(\sin\frac{\omega}{4}\Big)^{2L} e^{-i\frac{\omega}{2}L} u_0(\frac{\omega}{2})\\
&\quad + \Big(\sin\frac{\omega}{4}\Big)^{2L} e^{-i(\frac{\omega}{2}+\pi)L} |p_0(\frac{\omega}{2} + \pi)|^2 u(\frac{\omega}{2} + \pi);
\end{aligned}$$

that is, $T_p u(\omega)$ can be written as

$$T_p u(\omega) = \Big(\sin\frac{\omega}{4}\Big)^{2L} w_0(\frac{\omega}{2}),$$

where $w_0(\omega)$ is a trigonometric polynomial. This leads to

$$\frac{d^\ell}{d\omega^\ell}\Big(T_p u(\omega)\Big)\Big|_{\omega=0} = 0,$$

for all $0 \leq \ell < 2L$. Hence $T_p u(\omega) \in \mathbb{U}_{2L}$, as desired.　∎

For $v(\omega) \in \mathbb{V}_{2N+1}$, let $\|v\|_2$ be the L_2-norm of $v(\omega)$, defined by the square-root of the normalized inner product of v with itself, where the normalization is achieved

by dividing the integral over $[-\pi, \pi]$ by 2π, as defined in Sect. 4.1 and (4.1.2) in Sect. 6.1 of Chap. 6. Let A be a linear operator defined on a subspace $\mathbb{U}$ of $\mathbb{V}_{2N+1}$, and $\|\|A\|\|$ be the operator norm of A (see Definition 1 of Sect. 3.2 in Chap. 3 for matrix operators), namely:

$$\|\|A\|\| = \sup_{0 \neq u \in \mathbb{U}} \frac{\|Au\|_2}{\|u\|_2}.$$

Let $\sigma(A)$ denote the spectrum of A, i.e. the set of all eigenvalues of A, with **spectral radius** $\rho(A)$ defined to be the largest magnitude among all eigenvalues of A (see Theorem 1 in Sect. 3.2 of Chap. 3 for rectangular matrices). Then, by spectral theory (see Theorem 3 in Sect. 3.2 of Chap. 3), we have

$$\rho(A) = \lim_{n \to \infty} \|\|A^n\|\|^{\frac{1}{n}}.$$

In addition, the above (limit) formula is independent of the choice of the operator norm.

For $n \geq 1$, let D_n denote the union of two intervals defined by

$$D_n = [-2^n \pi, 2^n \pi] \setminus [-2^{n-1} \pi, 2^{n-1} \pi] = [-2^n \pi, -2^{n-1} \pi) \cup (2^{n-1} \pi, 2^n \pi].$$

Theorem 2 *Let $\phi \in L_2(\mathbb{R})$ be the refinable function associated with $p = \{p_k\}_{k=0}^N$. Suppose that p has sum-rule order L. Then, for an arbitrary given $\epsilon > 0$, there exists a positive constant c, such that for all $n \geq 1$,*

$$\int_{D_n} |\widehat{\phi}(\omega)|^2 d\omega \leq c \big(\rho(T_p|_{\mathbb{U}_{2L}}) + \epsilon\big)^n. \tag{10.4.2}$$

Proof Let

$$u_0(\omega) = (1 - \cos \omega)^L = \left(\sin \frac{\omega}{2}\right)^{2L}.$$

Then $u_0(\omega) \in \mathbb{U}_{2L}$. In addition,

$$u_0(\omega) \geq \left(\sin \frac{\pi}{4}\right)^{2L} = \frac{1}{2^L}, \text{ for } \omega \in D_1 = [-\pi, \pi] \setminus [-\frac{\pi}{2}, \frac{\pi}{2}].$$

Since $\widehat{\phi}(\omega)$ is continuous, and hence bounded, on $[-\pi, \pi]$, there exists a positive constant c_1, such that

$$|\widehat{\phi}(\omega)| \leq c_1, \text{ for } \omega \in [-\pi, \pi].$$

Then, with $\widehat{\phi}(\omega) = \prod_{j=1}^n p(\frac{\omega}{2^j})\widehat{\phi}(\frac{\omega}{2^n})$, we have

$$\int_{D_n} |\widehat{\phi}(\omega)|^2 d\omega = \int_{D_n} |\prod_{j=1}^{n} p(\frac{\omega}{2^j})|^2 \, |\widehat{\phi}(\frac{\omega}{2^n})|^2 d\omega$$

$$\leq c_1 \int_{D_n} |\prod_{j=1}^{n} p(\frac{\omega}{2^j})|^2 d\omega$$

$$\leq c_1 2^L \int_{D_n} |\prod_{j=1}^{n} p(\frac{\omega}{2^j})|^2 u_0(\frac{\omega}{2^n}) d\omega$$

$$\leq c_1 2^L \int_{-2^n \pi}^{2^n \pi} |\prod_{j=1}^{n} p(\frac{\omega}{2^j})|^2 u_0(\frac{\omega}{2^n}) d\omega$$

$$= c_1 2^L \int_{-\pi}^{\pi} (T_p^n u_0)(\omega) d\omega,$$

where the last equality follows from (9.1.12) in Sect. 9.1 of Chap. 9, on p.444. In addition, since

$$\rho(T_p\big|_{\mathbb{U}_{2L}}) = \lim_{n \to \infty} \||(T_p\big|_{\mathbb{U}_{2L}})^n\||^{\frac{1}{n}},$$

for the $\epsilon > 0$ given in the statement of the theorem, there exists a positive constant $c_\epsilon > 0$ such that, for $n \geq 1$,

$$\||(T_p\big|_{\mathbb{U}_{2L}})^n\|| \leq c_\epsilon \big(\rho(T_p\big|_{\mathbb{U}_{2L}}) + \epsilon\big)^n.$$

Therefore

$$\|T_p^n u_0\|_2 \leq \||(T_p\big|_{\mathbb{U}_{2L}})^n\|| \, \|u_0\|_2 \leq c_\epsilon \|u_0\|_2 \big(\rho(T_p\big|_{\mathbb{U}_{2L}}) + \epsilon\big)^n;$$

and we may conclude that

$$\int_{D_n} |\widehat{\phi}(\omega)|^2 d\omega \leq c_1 2^L \int_{-\pi}^{\pi} (T_p^n u_0)(\omega) d\omega$$

$$\leq c_1 2^L \sqrt{2\pi} \|T_p^n u_0\|_2 \leq c_1 2^L \sqrt{2\pi} c_\epsilon \|u_0\|_2 \big(\rho(T_p\big|_{\mathbb{U}_{2L}}) + \epsilon\big)^n.$$

This proves that (10.4.2) holds, with the choice of $c = c_1 2^L \sqrt{2\pi} c_\epsilon \|u_0\|_2$.   ∎

Next we give a Sobolev smoothness estimate for ϕ in terms of the spectral radius $\rho(T_p\big|_{\mathbb{U}_{2L}})$. Throughout this section, log denotes the common logarithm; that is, using base 10.

Theorem 3 | Sobolev smoothness estimate | *Let $\phi \in L_2(\mathbb{R})$ be the refinable function associated with $p = \{p_k\}_{k=0}^{N}$. Suppose that p has sum-rule order L. Then*

$$\nu_\phi \geq \gamma_0 = -\frac{\log \rho(T_p\big|_{\mathbb{U}_{2L}})}{2\log 2}. \tag{10.4.3}$$

In addition, if ϕ is stable, then $\nu_\phi = \gamma_0$.

Proof Here, we only give the proof of (10.4.3), but remark that the validity of the second statement ($\nu_\phi = \gamma_0$) depends on the fact that L is the largest integer (that is, the order) for which the sum rules hold. For any $\gamma < \gamma_0$, there exists an $\epsilon > 0$, such that

$$\gamma < -\frac{\log\left(\rho(T_p\big|_{\mathbb{U}_{2L}}) + \epsilon\right)}{2\log 2},$$

or equivalently,

$$4^\gamma\left(\rho(T_p\big|_{\mathbb{U}_{2L}}) + \epsilon\right) < 1.$$

From Theorem 2, we have

$$\int_{-\infty}^{\infty} (1 + |\omega|^2)^\gamma |\widehat{\phi}(\omega)|^2 d\omega$$

$$= \int_{-\pi}^{\pi} (1 + |\omega|^2)^\gamma |\widehat{\phi}(\omega)|^2 d\omega + \sum_{n=1}^{\infty} \int_{D_n} (1 + |\omega|^2)^\gamma |\widehat{\phi}(\omega)|^2 d\omega$$

$$\leq c_2 + \sum_{n=1}^{\infty} \left(1 + (2^n\pi)^2\right)^\gamma \int_{D_n} |\widehat{\phi}(\omega)|^2 d\omega$$

$$\leq c_2 + \sum_{n=1}^{\infty} \left(1 + (2^n\pi)^2\right)^\gamma c\left(\rho(T_p\big|_{\mathbb{U}_{2L}}) + \epsilon\right)^n$$

$$< \infty,$$

since $4^\gamma\left(\rho(T_p\big|_{\mathbb{U}_{2L}}) + \epsilon\right) < 1$, where c_2 is some positive constant. This shows that $\phi \in \mathbb{W}^\gamma(\mathbb{R})$. Hence, (10.4.3) holds. ∎

Next, we will introduce the notion of generalized eigenvectors (in Theorem 4) in order to derive a relatively simple method for computing the spectral radius $\rho(T_p\big|_{\mathbb{U}_{2L}})$. The formula for this computation will be established in Theorem 5. For an integer $\ell \geq 0$, let $\mathbf{s}_\ell$ denote the $1 \times (2N + 1)$ row vectors

$$\mathbf{s}_\ell = \left[k^\ell\right]_{-N \leq k \leq N} = \left[(-N)^\ell, \ldots (-1)^\ell, 0^\ell, 1^\ell, \ldots, N^\ell\right], \tag{10.4.4}$$

where the standard convention $0^0 = 1$ is adopted.

For the trigonometric polynomial $v(\omega) = \sum_{k=-N}^{N} v_k e^{-ik\omega} \in \mathbb{V}_{2N+1}$, let $\text{vec}(v)$ denote the column vector defined by (9.1.9) on p.442, namely:

$$\text{vec}(v) = [v_{-N}, \ldots, v_{-1}, v_0, v_1, \ldots, v_N]^T.$$

Observe that

$$\mathbf{s}_\ell \mathrm{vec}(v) = \sum_{k=-N}^{N} k^\ell v_k = i^\ell \frac{d^\ell}{d\omega^\ell} v(0). \qquad (10.4.5)$$

Thus, we may conclude that $u(\omega) \in \mathbb{U}_{2L}$, if and only if

$$\mathbf{s}_\ell \mathrm{vec}(u) = 0, \text{ for all } 0 \le \ell < 2L.$$

Theorem 4 $\boxed{\textbf{Generalized } \dfrac{1}{2^\ell}\textbf{-eigenvectors of } \mathcal{T}_p}$ *Suppose that* $p = \{p_k\}_{k=0}^{N}$ *has sum-rule order* L. *Then* $\mathbf{s}_\ell, 0 \le \ell < 2L$, *are generalized (left) eigenvectors of* $\mathcal{T}_p$ *associated with eigenvalues* $\frac{1}{2^\ell}$, *in the sense that*

$$\mathbf{s}_\ell \mathcal{T}_p = \frac{1}{2^\ell}\mathbf{s}_\ell + \sum_{j=0}^{\ell-1} c_{\ell,j}\, \mathbf{s}_j, \qquad (10.4.6)$$

for some $c_{\ell,j} \in \mathbb{C}$.

When $\ell = 0$, (10.4.6) is understood to be $\mathbf{s}_0 \mathcal{T}_p = \mathbf{s}_0$, which can be shown directly from the definition of $\mathcal{T}_p$ and the fact that p has at least sum-rule order 1; that is,

$$\sum_k p_{2k} = \sum_k p_{2k+1} = 1.$$

$\boxed{\textbf{Proof of Theorem 4}}$ To prove the theorem, we start by noting that

$$\frac{d^\ell}{d\omega^\ell}\left(|p(\frac{\omega}{2} + \pi)|^2\right)\Big|_{\omega=0} = 0, \ 0 \le \ell < 2L. \qquad (10.4.7)$$

This can be obtained from the assumption that p has sum-rule order L, if and only if its two-scale symbol $p(\omega)$ satisfies $\frac{d^\ell}{d\omega^\ell} p(\pi) = 0$ for $0 \le \ell < L$. We leave the verification of (10.4.7) as an exercise.

To prove (10.4.6), it is sufficient to prove that, for any $v(\omega) \in \mathbb{V}_{2N+1}$,

$$\mathbf{s}_\ell \mathcal{T}_p \mathrm{vec}(v) = \frac{1}{2^\ell}\mathbf{s}_\ell \mathrm{vec}(v) + \sum_{j=0}^{\ell-1} c_{\ell,j}\, \mathbf{s}_j \mathrm{vec}(v). \qquad (10.4.8)$$

Since $\mathcal{T}_p \mathrm{vec}(v) = \mathrm{vec}(T_p v)$ and from (10.4.5), we have

$$\mathbf{s}_\ell \, \mathcal{T}_p \text{vec}(v) = \mathbf{s}_\ell \, \text{vec}(T_p v) = i^\ell \frac{d^\ell}{d\omega^\ell} \Big(T_p v(\omega) \Big)\Big|_{\omega=0}$$

$$= i^\ell \frac{d^\ell}{d\omega^\ell} \Big(|p(\tfrac{\omega}{2})|^2 v(\tfrac{\omega}{2}) + |p(\tfrac{\omega}{2} + \pi)|^2 v(\tfrac{\omega}{2} + \pi) \Big)\Big|_{\omega=0}$$

$$= i^\ell \frac{d^\ell}{d\omega^\ell} \Big(|p(\tfrac{\omega}{2})|^2 v(\tfrac{\omega}{2}) \Big)\Big|_{\omega=0} \text{(by(10.4.7))}$$

$$= i^\ell \sum_{j=0}^{\ell} \binom{\ell}{j} \frac{d^{\ell-j}}{d\omega^{\ell-j}} \Big(|p(\tfrac{\omega}{2})|^2 \Big)\Big|_{\omega=0} \frac{1}{2^j} \frac{d^j}{d\omega^j} v(0)$$

$$= i^\ell p(0) \frac{1}{2^\ell} \frac{d^\ell}{d\omega^\ell} v(0) + \sum_{j=0}^{\ell-1} c_{\ell,j} i^j \frac{d^j}{d\omega^j} v(0)$$

$$= \frac{1}{2^\ell} \mathbf{s}_\ell \text{vec}(v) + \sum_{j=0}^{\ell-1} c_{\ell,j} \, \mathbf{s}_j \text{vec}(v),$$

where

$$c_{\ell,j} = i^{\ell-j} \binom{\ell}{j} \frac{1}{2^j} \frac{d^{\ell-j}}{d\omega^{\ell-j}} \Big(|p(\tfrac{\omega}{2})|^2 \Big)\Big|_{\omega=0}.$$

This shows that (10.4.8) holds, and we may now conclude that $\mathbf{s}_\ell, 0 \le \ell < 2L$, are generalized $\frac{1}{2^\ell}$-eigenvectors of $\mathcal{T}_p$. ∎

Example 1 Let ϕ be the hat function, which is refinable with the refinement mask $p = \{p_k\}_{k=0}^2$, where $p_0 = p_2 = \frac{1}{2}$, $p_1 = 1$. In Example 1 of Sect. 9.1 on p.439, it was shown that the representation matrix $\mathcal{T}_p$ of the transition operator T_p (restricted to $\mathbb{V}_5$) associated with p is given by

$$\mathcal{T}_p = \frac{1}{8} \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 6 & 4 & 1 & 0 & 0 \\ 1 & 4 & 6 & 4 & 1 \\ 0 & 0 & 1 & 4 & 6 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

Since p has sum-rule order $L = 2$ and the eigenvalues of $\mathcal{T}_p$ are given by $1, \frac{1}{2}, \frac{1}{2^2}, \frac{1}{2^3}$ (see Example 2 in Sect. 9.1 on p.443), we may apply Theorem 4 to compute the corresponding generalized eigenvectors $\mathbf{s}_\ell, 0 \le \ell < 2L = 4$, given by

$$\mathbf{s}_0 = [1, 1, 1, 1, 1],$$
$$\mathbf{s}_1 = [-2, -1, 0, 1, 2],$$
$$\mathbf{s}_2 = [(-2)^2, (-1)^2, 0, 1^2, 4^2] = [4, 1, 0, 1, 4],$$
$$\mathbf{s}_3 = [(-2)^3, (-1)^3, 0, 1^3, 2^3] = [-8, -1, 0, 1, 8].$$

To verify that they are indeed generalized eigenvectors of $\mathcal{T}_p$, we may directly compute

$$\mathbf{s}_0 \mathcal{T}_p = [1, 1, 1, 1, 1] = \mathbf{s}_0,$$

$$\mathbf{s}_1 \mathcal{T}_p = [-1, -\frac{1}{2}, 0, \frac{1}{2}, 1] = \frac{1}{2}\mathbf{s}_1,$$

$$\mathbf{s}_2 \mathcal{T}_p = \frac{1}{4}[5, 2, 1, 2, 5] = \frac{1}{4}\mathbf{s}_2 + \frac{1}{4}\mathbf{s}_0,$$

$$\mathbf{s}_3 \mathcal{T}_p = \frac{1}{4}[-7, -2, 0, 2, 7] = \frac{1}{8}\mathbf{s}_3 + \frac{3}{8}\mathbf{s}_1.$$

∎

Computation of $\rho(T_p|_{\mathbb{U}_{2L}})$

Let S_{2L} denote the subspace of $\mathbb{C}^{2N+1}$ defined by

$$S_{2L} = \text{span}\{\mathbf{s}_\ell : 0 \le \ell < 2L\},$$

where $\mathbf{s}_\ell, 0 \le \ell < 2L$, are defined by (10.4.4). From Theorem 4, we know that S_{2L} is invariant under $\mathcal{T}_p$, and

$$\sigma(\mathcal{T}_p|_{S_{2L}}) = \{1, \frac{1}{2}, \dots, \frac{1}{2^{2L-1}}\}. \tag{10.4.9}$$

Let $S_{2L}^\perp$ be the orthogonal complement of S_{2L} in $\mathbb{C}^{2N+1}$:

$$\begin{aligned} S_{2L}^\perp &= \{\mathbf{v} \in \mathbb{C}^{2N+1} : \mathbf{s}_\ell \cdot \mathbf{v} = 0, \text{ for all } 0 \le \ell < 2L\} \\ &= \{\mathbf{v} \in \mathbb{C}^{2N+1} : \sum_{k=-N}^{N} k^\ell v_k = 0, \text{ for all } 0 \le \ell < 2L\}. \end{aligned}$$

Recall that $\mathbb{U}_{2L}$, defined by (10.4.1), can be written as

$$\mathbb{U}_{2L} = \left\{ u(\omega) = \sum_{k=-N}^{N} u_k e^{-i\omega k} : \sum_{k=-N}^{N} k^\ell u_k = 0, \text{ for all } 0 \le \ell < 2L \right\}.$$

Thus, we have

$$u(\omega) \in \mathbb{U}_{2L} \Rightarrow \text{vec}(u) \in S_{2L}^\perp, \quad \text{and}$$

$$\mathbf{u} = [u_{-N}, \dots, u_0, \dots, u_N]^T \in S_{2L}^\perp \Rightarrow u(\omega) = \sum_{k=-N}^{N} u_k e^{-ik\omega} \in \mathbb{U}_{2L}.$$

Hence,

$$\sigma(T_p|_{\mathbb{U}_{2L}}) = \sigma(\mathcal{T}_p|_{S_{2L}^\perp}). \tag{10.4.10}$$

On the other hand, from $\mathbb{C}^{2N+1} = S_{2L} \oplus^{\perp} S_{2L}^{\perp}$, we obtain

$$\sigma(\mathcal{T}_p) = \sigma(\mathcal{T}_p\big|_{S_{2L}}) \cup \sigma(\mathcal{T}_p\big|_{S_{2L}^{\perp}}). \tag{10.4.11}$$

From (10.4.9)–(10.4.11), we have

$$\sigma(T_p\big|_{\mathbb{U}_{2L}}) = \sigma(\mathcal{T}_p)\backslash\sigma(\mathcal{T}_p\big|_{S_{2L}}) = \sigma(\mathcal{T}_p)\backslash\{1, \frac{1}{2}, \ldots, \frac{1}{2^{2L-1}}\}.$$

Hence, $\rho(T_p\big|_{\mathbb{U}_{2L}})$ is given by

$$\boxed{\rho_0 = max\left\{|\lambda| : \lambda \in \sigma(\mathcal{T}_p)\backslash\{1, \frac{1}{2}, \ldots, \frac{1}{2^{2L-1}}\}\right\}}. \tag{10.4.12}$$

This, together with Theorem 3, leads to the following result.   ■

Theorem 5 $\boxed{\textbf{Smoothness of } \phi \textbf{ in terms of eigenvalues of } \mathcal{T}_p}$ *Let $\phi \in L_2(\mathbb{R})$ be the refinable function associated with a refinement mask $p = \{p_k\}_{k=0}^{N}$ that has sum-rule order L. Then*

$$\nu_\phi \geq -\frac{\log \rho_0}{2 \log 2}, \tag{10.4.13}$$

where ρ_0 is defined by (10.4.12). In addition, if ϕ is stable, then $\nu_\phi = -\frac{\log \rho_0}{2 \log 2}$.

Theorem 5 provides a simple formula for computing the Sobolev regularity (that is, smoothness) of refinable functions (and hence, of wavelets). Essentially, this regularity is determined by the largest magnitude (i.e. absolute values) among the eigenvalues of $\mathcal{T}_p$, after excluding the special eigenvalues $1, 1/2, \ldots, 1/2^{(2L-1)}$. The number $-\frac{\log \rho_0}{2 \log 2}$ in (10.4.13) is called a (Sobolev) smoothness estimate for ϕ. If ϕ is stable and p has (exact) sum-rule order L, then this estimate is optimal.

Example 2 Let ϕ be the hat function. Recall from Example 2 in Sect. 9.1 of Chap. 9 on p.443 that the eigenvalues of $\mathcal{T}_p$ are $1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}$. Also, recall that the refinement mask p for ϕ has (exact) sum-rule order 2. Since after excluding $1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}$, the only remaining eigenvalue of $\mathcal{T}_p$ is $\frac{1}{8}$, we may conclude that $\rho_0 = \frac{1}{8}$. In addition, since ϕ is stable, we have

$$\nu_\phi = -\frac{\log(\rho_0)}{2 \log 2} = -\frac{\log(\frac{1}{8})}{2 \log 2} = \frac{3}{2}.$$

We may verify the optimality of the Sobolev regularity, $-\frac{\log(\frac{1}{8})}{2 \log 2}$, by directly calculating ν_ϕ from ϕ. Indeed, with

$$\widehat{\phi}(\omega) = (\chi_{[0,1)} * \chi_{[0,1)})^{\wedge}(\omega) = (\frac{e^{-i\omega} - 1}{-i\omega})^2 = 2e^{-i\omega}\frac{1 - \cos \omega}{\omega^2},$$

we have

$$\int_{-\infty}^{\infty} |\widehat{\phi}(\omega)|^2 (1+\omega^2)^\gamma d\omega = 4 \int_{-\infty}^{\infty} \frac{(1-\cos\omega)^2}{\omega^4} (1+\omega^2)^\gamma d\omega$$

$$= 8 \int_{0}^{\infty} \frac{(1-\cos\omega)^2}{\omega^4} (1+\omega^2)^\gamma d\omega$$

$$= 8 \int_{0}^{1} \frac{(1-\cos\omega)^2}{\omega^4} (1+\omega^2)^\gamma d\omega$$

$$+ 8 \int_{1}^{\infty} \left(\frac{3}{2} - 2\cos\omega + \frac{1}{2}\cos 2\omega \right) \frac{(1+\omega^2)^\gamma}{\omega^4} d\omega.$$

Observe that the first integral is finite (since for any $\gamma \geq 0$, $\frac{(1-\cos\omega)^2}{\omega^4}(1+\omega^2)^\gamma$ is bounded on $[0, 1]$), and that the second integral is convergent, if and only if $4-2\gamma > 1$ (i.e. $\gamma < \frac{3}{2}$). Thus $\phi \in \mathbb{W}^\gamma(\mathbb{R})$ for any $\gamma < \frac{3}{2}$, but $\phi \notin \mathbb{W}^{\frac{3}{2}}(\mathbb{R})$. Hence, $\nu_\phi = \frac{3}{2}$, which verifies that $-\frac{\log(\rho_0)}{2\log 2}$ is the optimal Sobolev estimate.   ∎

Example 3 Let $\phi(x) = \frac{1}{3}\chi_{[0,3)}(x)$ be the refinable function with the two-scale symbol

$$p(\omega) = \frac{1}{2}(1 + e^{-i3\omega}).$$

In Example 5 of Sect. on p.526, we showed that the spectrum (the set of eigenvalues) of $\mathcal{T}_p$ is given by

$$\sigma(\mathcal{T}_p) = \{1, 1, -1, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, -\frac{1}{2}\}.$$

It is clear that the refinement mask has (exact) sum-rule order 1. Thus,

$$\sigma(\mathcal{T}_p) \setminus \{1, \frac{1}{2}\} = \{1, -1, \frac{1}{2}, \frac{1}{2}, -\frac{1}{2}\}.$$

Hence, $\rho_0 = 1$, and the smoothness estimate in (10.4.13) is given by

$$\nu_\phi \geq -\frac{\log \rho_0}{2\log 2} = 0.$$

One can also show directly from $\widehat{\phi}(\omega)$ that $\nu_\phi = \frac{1}{2}$ (this is left as an exercise). Hence, (10.4.13) does not provide a good smoothness estimate for ϕ. Of course, the reason is that ϕ is not stable.   ∎

Example 4 Let ϕ be the D_4 scaling function. It was shown in Example 2 of Sect. on p.513 that the eigenvalues of $\mathcal{T}_p$ are given by

$$1, \frac{1}{2}, \frac{1}{4}, \frac{1}{4}, \frac{1}{8}, -\frac{1}{16}, -\frac{1}{16}.$$

The associated refinement mask p has (exact) sum-rule order 2. Thus, we have

$$\sigma(\mathcal{T}_p) \backslash \{1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}\} = \{\frac{1}{4}, -\frac{1}{16}\},$$

so that $\rho_0 = \frac{1}{4}$. Since ϕ is orthogonal (and hence, stable), it follows from Theorem 5 that

$$\nu_\phi = -\frac{\log(\rho_0)}{2 \log 2} = -\frac{\log(\frac{1}{4})}{2 \log 2} = 1.$$

∎

Example 5   Let $\widetilde{\phi}$ be the scaling function associated with one of the 5/3-tap biorthogonal lowpass filters given by

$$\widetilde{p}(\omega) = -\frac{1}{8} e^{i\omega} + \frac{1}{4} + \frac{3}{4} e^{-i\omega} + \frac{1}{4} e^{-i2\omega} - \frac{1}{8} e^{-i3\omega}.$$

The associated refinement mask $\widetilde{p}$ has (exact) sum-rule order 2. From Example 7 of Sect. 10.3 on p.530, we know that the largest magnitude among the eigenvalues of $\mathcal{T}_p$, excluding $1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}$, is 0.54279. Since $\widetilde{\phi}$ is stable, it follows from Theorem 5 that

$$\nu_{\widetilde{\phi}} = -\frac{\log(\rho_0)}{2 \log 2} = -\frac{\log(0.54279)}{2 \log 2} = 0.44077.$$

∎

In the literature, there are other formulas similar to the one given in Theorem 5 for estimating Sobolev regularities (smoothness) for compactly supported wavelets of two or more variables, and for compactly supported multi-wavelets (that is, vector-valued wavelets) of one variable or several variables. For a given refinement mask, it is not difficult to write a program to calculate these Sobolev smoothness estimates. For example, Matlab programs for computing smoothness estimates of compactly supported wavelets and multi-wavelets can be downloaded from `www.math.umsl.edu/~jiang` .

Exercises

Exercise 1   Suppose $p = \{p_k\}_{k=0}^N$ has at least sum-rule order 1. Show directly that $\mathbf{s}_0 = [1, \ldots, 1]$ is a (left) 1-eigenvector of $\mathcal{T}_p$.

Exercise 2   Derive (10.4.7).

Exercise 3   Let $\phi(x) = \chi_{[0,1)}(x)$ be the Haar scaling function. Apply the smoothness formula in Theorem 5 to compute ν_ϕ. Then verify directly from $\widehat{\phi}(\omega)$ that ν_ϕ is the optimal smoothness estimate.

Exercise 4 Prove (10.4.11).

Hint: Show that $\mathcal{T}_p$ is similar to a block matrix.

Exercise 5 Let $\phi(x) = \frac{1}{3}\chi_{[0,3)}(x)$. Show that $\phi(x) \in \mathbb{W}^\gamma(\mathbb{R})$ for any $\gamma < \frac{1}{2}$, but $\phi(x) \notin \mathbb{W}^{\frac{1}{2}}(\mathbb{R})$, and hence conclude that $\nu_\phi = \frac{1}{2}$.

Exercise 6 Let $\phi(x) = \frac{1}{2}\chi_{[0,2)}(x)$. Compute the critical Sobolev exponent ν_ϕ of ϕ.

Exercise 7 Let $p = \{p_k\}_{k=0}^3$ be a refinement mask with

$$p_0 = p_3 = \frac{1}{4}, \, p_1 = p_2 = \frac{3}{4}.$$

Let $\mathbf{s}_\ell$, $\ell = 0, 1, 2, 3$, be the vectors defined by (10.4.4) with $N = 3$. Verify that $\mathbf{s}_\ell$ are generalized eigenvectors of $\mathcal{T}_p$ by finding $c_{\ell,j}$ to satisfy (10.4.6). (For reference, see Exercise 4 on p.444.)

Exercise 8 Let ϕ be the D_4 scaling function, and $\mathcal{T}_p$ be the representation matrix of the transition operator given in Example 2 of Sect. 10.2 on p.513. Let $\mathbf{s}_\ell$, $\ell = 0, 1, 2, 3$, be the vectors defined by (10.4.4) with $N = 3$. Verify that $\mathbf{s}_\ell$ are generalized eigenvectors of $\mathcal{T}_p$ by finding $c_{\ell,j}$ to satisfy (10.4.6).

Exercise 9 Let $p = \{p_k\}_{k=0}^2$ with

$$p_0 = \frac{3}{4}, \, p_1 = 1, \, p_2 = \frac{1}{4},$$

be the refinement mask considered in Example 1 in Sect. 10.2 on p.511. Apply Theorem 5 to compute the critical Sobolev exponent ν_ϕ of the associated refinable function ϕ.

Exercise 10 Repeat Exercise 9 for $p = \{p_k\}_{k=0}^2$ with

$$p_0 = \frac{2}{3}, \, p_1 = 1, \, p_2 = \frac{1}{3}.$$

(For reference, see Exercise 2 on p.516.)

Exercise 11 (Computation by Matlab) Let ϕ and $\widetilde{\phi}$ denote the scaling functions associated with the 9/7-biorthogonal filters. Apply Theorem 5 to calculate their critical Sobolev exponents ν_ϕ and $\nu_{\widetilde{\phi}}$.

Index